# SCIENCE AND TECHNOLOGY TEXT MINING: MEXICO CORE COMPETENCIES

By

Dr. Ronald N. Kostoff
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22304
Phone: 703-696-4198
Fax: 703-696-4274
Internet: kostofr@onr.navy.mil

Dr. J. Antonio del Río and Héctor D. Cortés
Centro de Investigación en Energía, UNAM
Temixco, Mor. México

Mr. Charles Smith
Booz-Allen Hamilton
Bethesda, MD

Dr. Andrew Smith
University of Queensland
Australia

Ms. Caroline Wagner
University of Amsterdam
Amsterdam, the Netherlands

Dr. Loet Leydesdorff
University of Amsterdam
Amsterdam, The Netherlands

Dr. George Karypis
University of Minnesota
Minneapolis, MN 55455

Mr. Guido Malpohl
University of Karlsruhe
Postfach 6980
76128 Karlsruhe, Germany
Internet: malpohl@ipd.uka.de

Mr. Rene Tshiteya
DDL-OMNI Engineering, LLC,
8260 Greensboro Drive, Suite 600,

| | |
|---|---|
| # Report Documentation Page | *Form Approved*<br>*OMB No. 0704-0188* |

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

| | | |
|---|---|---|
| 1. REPORT DATE<br>**2002** | 2. REPORT TYPE<br>**N/A** | 3. DATES COVERED<br>**-** |

| | |
|---|---|
| 4. TITLE AND SUBTITLE<br>**Science and Technology Text Mining: Mexico Core Competencies** | 5a. CONTRACT NUMBER |
| | 5b. GRANT NUMBER |
| | 5c. PROGRAM ELEMENT NUMBER |
| 6. AUTHOR(S) | 5d. PROJECT NUMBER |
| | 5e. TASK NUMBER |
| | 5f. WORK UNIT NUMBER |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)<br>**Office of Naval Research Dr. Ronald N. Kostoff 800 N. Quincy Street Arlington, VA 22304** | 8. PERFORMING ORGANIZATION REPORT NUMBER |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | 10. SPONSOR/MONITOR’S ACRONYM(S) |
| | 11. SPONSOR/MONITOR’S REPORT NUMBER(S) |

12. DISTRIBUTION/AVAILABILITY STATEMENT
**Approved for public release, distribution unlimited**

13. SUPPLEMENTARY NOTES
**The original document contains color images.**

14. ABSTRACT

15. SUBJECT TERMS

| | | | | | |
|---|---|---|---|---|---|
| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT<br>**UU** | 18. NUMBER OF PAGES<br>**117** | 19a. NAME OF RESPONSIBLE PERSON |
| a. REPORT<br>**unclassified** | b. ABSTRACT<br>**unclassified** | c. THIS PAGE<br>**unclassified** | | | |

Mclean, VA 22102

(*The views in this paper are solely those of the authors, and do not necessarily represent the views of the Department of the Navy or any of its components, the UNAM, Booz-Allen Hamilton, DDL-OMNI, the University of Queensland, the University of Amsterdam, the University of Karlsruhe, or the University of Minnesota*.)

**KEYWORDS:** Mexico; Science and Technology; Bibliometrics; Computational Linguistics; Core Competencies; Research Evaluation; Factor Analysis; Concept Clustering; Document Clustering; Data Compression; Network Analysis; Leximancer; CLUTO; Greedy String Tiling

## 1) ABSTRACT

The structure and infrastructure of the Mexican technical literature was determined. A representative database of technical articles was extracted from the Science Citation Index for the year 2002, with each article containing at least one author with a Mexican address. Many different manual and statistical clustering methods were used to identify the structure of the technical literature (especially the science and technology core competencies), and to evaluate the strengths and weaknesses of each technique. One of the pervasive technical topics identified from the clustering, Thin Films research, was analyzed further using bibliometrics, in order to identify the infrastructure of this technology.

## 2) BACKGROUND

Country Technology Assessments

National science and technology (S&T) core competencies represent a country's strategic capabilities in S&T. Knowledge of country core competencies is important for myriad reasons:

a) Priority technical areas for joint commercial or military ventures
b) Assessment of a country's military potential
c) Knowledge of emerging areas to avoid commercial or military surprise

Obtaining such global technical awareness, especially from the literature, is difficult for multiple reasons:

a) Much science and technology performed is not documented
b) Much documented science and technology is not widely available
c) Much available documented science and technology is expensive and difficult to acquire
d) Few credible techniques exist for extracting useful information from large amounts of science and technology documentation (1)

Most credible country technology assessments are based on a combination of personal visitations to the country of interest, supplemented by copious reading of technology reports from that country. Such processes tend to be laborious, slow, expensive, and accompanied by large gaps in the knowledge available. The more credible and complete evaluation processes will focus on selected technologies from a particular country, and provide in-depth analysis.

For the past half century, driven mainly by the Cold War, a large number of country technology assessments were performed (2-14). The last decade has seen an expansion in focus to technologies of major economic competitors. Over the past two decades, some of the most credible of these country technology assessments have come from two organizations: World Technology Evaluation Center (WTEC-Loyola Univ) and Foreign Applied Sciences Assessment Center (FASAC-SAIC). In conducting their studies, both of these organizations would gather topical literature from the country of interest, assemble teams of experts in the topical area, have the teams review the literature as well as conduct site visitations, and have the teams brief their findings and write a final report. The studies performed by these groups remain seminal approaches to country technology assessments.

Text Mining Technology Assessments

The first author's group has been developing text mining approaches to extract useful information from the global science and technology literature for the past decade (15-26).

These studies have typically focused on a technical discipline, and have examined global S&T efforts in this discipline.  It is believed that such approaches, with slight modification, could be adapted to identifying the core S&T competencies in selected countries or regions, including estimation of the relative levels of effort in each of the core technology areas.  It is also believed that coupling of the text mining approach with WTEC and FASAC approaches would amplify the strengths of each approach and reduce the limitations.  The text mining component would be performed initially to identify:

- Key core competencies and technology thrusts in the country of interest
- Key interdisciplinary thrusts
- Approximate levels of efforts in technology-specific competency areas and in interdisciplinary areas
- Highly productive researchers
- Highly productive Centers of Excellence, including those not well known
- Highly cited researchers

Once the key technologies, researchers, and Centers of Excellence had been identified, then site visitation strategies could be developed.  The second phase of the effort would be the actual site visitations.  A key step in this hybrid process would be demonstration of the ability of text mining to identify the targets of interest with reasonable precision in a timely manner at an acceptable cost.  These three driving parameters (performance, time, cost) could be traded-off against each other to provide  a balance acceptable and tailored to a variety of potential customers.

## Mexican Science and Technology Structure

Mexico is an important country with which our current President and Administration want to strengthen relationships and build a stronger partnership.  In addition, there is a long common border, with common security concerns.  To improve awareness of Mexico’s S&T program, Mexico was selected as the prototype for a country core competency assessment.

This sub-section provides background information about the sponsorship of research in Mexico.  It describes the Federal Sciences and Technology Expenditures (FSTE), the Gross Domestic Expenditures in Science and the number of researchers involved in scientific activities according to Mexican government data.

### *Mexican Budget for Science and Technology*

This section summarizes the Mexican budget for S&T.  The Federal Mexican S&T expenditures (FSTE ) have been almost constant during the last decade, oscillating around 0.4% of the Gross Domestic Product (GDP).  In terms of the Discretionary Federal Budget (DFB), the FSTE ratio has been of the order of 2.5%. This is the lowest FSTE in the thirty member nations of the Organization for Economic Cooperation and Development (OCDE).

Evolution of the Federal S&T expenditures of Mexico from 1992 to 2002 can be discerned from Table 1.

TABLE 1

| Federal science and technology expenditures (FSTE) | | | | | |
|---|---|---|---|---|---|
| Year | FSTE Millions of constant pesos 2000 | Gross Domestic Product Millions of constant pesos 2000 | FSTE/GDP | Discretionary Federal Budget Millions of constant pesos 2000 | FSTE/DFB |
| 1992 | 15102 | 4704002 | 0.32% | 745169 | 2.03% |
| 1993 | 17514 | 4795755 | 0.37% | 790211 | 2.22% |
| 1994 | 20332 | 5007503 | 0.41% | 879672 | 2.31% |
| 1995 | 16584 | 4698691 | 0.35% | 742840 | 2.23% |
| 1996 | 17293 | 4940829 | 0.35% | 789276 | 2.19% |
| 1997 | 22236 | 5275421 | 0.42% | 877705 | 2.53% |
| 1998 | 25626 | 5540794 | 0.46% | 865160 | 2.96% |
| 1999 | 23483 | 5741525 | 0.41% | 888945 | 2.64% |
| 2000 | 25586 | 6122609 | 0.42% | 965162 | 2.65% |
| 2001 | 25373 | 6103831 | 0.42% | 991119 | 2.56% |
| 2002 | 25374 | 6152829 | 0.41% | 1026820 | 2.47% |

*Mexican Gross Domestic Expenditures on Research*

The evolution of the Distribution of the Gross Domestic Expenditures on Research (GDERD) and Development in terms of the funding sector and the final area is shown in Table 2. In this table, it can be seen that the Government and the Universities are allocated most of the GDERD, mainly in the natural sciences and engineering areas.

TABLE 2

Gross Domestic Expeditures on Research and Development
Thousand of constant pesos 2002

| Executive sector | Area or research | 1993 | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 |
|---|---|---|---|---|---|---|---|---|---|---|
| Business enterprise | Natural sciences and Engineering. | 1,092,574 | 3,564,332 | 2,962,705 | 3,398,742 | 3,573,523 | 5,654,908 | 5,890,104 | 6,327,096 | 6,904,117 |
| | Social sciences and humanities. | 0 | 146,171 | 56,000 | 24,724 | 15,453 | 239,831 | 412,964 | 478,144 | 437,595 |
| **Total Business enterprise** | | **1,092,574** | **3,710,503** | **3,018,705** | **3,423,466** | **3,588,976** | **5,894,739** | **6,303,069** | **6,805,240** | **7,341,712** |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Government | Natural sciences and Engineering. | 3,278,326 | 3,535,056 | 4,288,193 | 4,937,675 | 6,238,865 | N.A. | N.A. | 7,418,760 | 7,325,696 |
| | Social sciences and humanities. | 465,318 | 535,672 | 513,129 | 639,889 | 809,245 | N.A. | N.A. | 2,122,470 | 2,142,143 |
| **Total government** | | **3,743,644** | **4,070,728** | **4,801,322** | **5,577,564** | **7,048,111** | **7,696,800** | **11,105,414** | **9,541,230** | **9,467,839** |
| **Higher education** | Natural sciences and Engineering. | 4,429,666 | 4,981,674 | 4,877,776 | 4,537,054 | 5,681,877 | 5,014,247 | 4,632,961 | 4,534,679 | 5,075,542 |
| | Social sciences and humanities. | 1,242,247 | 1,881,326 | 1,787,430 | 1,266,041 | 1,572,705 | 1,587,090 | 1,868,986 | 1,931,589 | 2,295,732 |
| **Total higher education** | | **5,671,913** | **6,863,000** | **6,665,207** | **5,803,094** | **7,254,582** | **6,601,337** | **6,501,947** | **6,466,268** | **7,371,274** |
| **Private non profit** | Natural sciences and Engineering. | 18,844 | 27,263 | 30,471 | 103,272 | 54,020 | 629,245 | 662,071 | 18,554 | 14,696 |
| | Social sciences and humanities. | 27,839 | 27,281 | 31,031 | 408,197 | 243,182 | 100,994 | 107,584 | 40,899 | 40,082 |
| **Total private non profit** | | **46,683** | **54,544** | **61,502** | **511,469** | **297,202** | **730,239** | **769,654** | **59,453** | **54,778** |
| **Total** | Natural sciences and Engineering. | 8,819,411 | 12,108,325 | 12,159,145 | 12,976,742 | 15,548,286 | N.A. | N.A. | 18,299,089 | 19,320,051 |
| | Social sciences and humanities. | 1,735,403 | 2,590,450 | 2,387,591 | 2,338,851 | 2,640,586 | N.A. | N.A. | 4,573,102 | 4,915,552 |
| **TOTAL GERD** | | **10,554,815** | **14,698,775** | **14,546,735** | **15,315,593** | **18,188,871** | **20,923,116** | **24,680,084** | **22,872,191** | **24,235,603** |

In Table 3, the evolution of the GDERD invested in basic, applied and experimental development illustrates that government expenditures in basic and applied research are similar, while business expenditures are larger in applied research. Total expenditures are concentrated in government and education institutions.

TABLE 3

Gross Domestic Expeditures on Research and Development
Thousand of constant pesos 2002

| Executive sector | Area or research | 1993 | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 |
|---|---|---|---|---|---|---|---|---|---|---|
| Business enterprise | Basic research. Applied | 73,307 | 166,749 | 173,013 | 142,094 | 192,482 | N.A. | N.A. | 493,261 | 556,978 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | research. | 198,321 | 1,409,091 | 793,375 | 1,413,534 | 1,516,127 | 399,205 | 450,526 | 2,688,203 | 2,645,532 |
| | Experimental development | 604,606 | 1,679,152 | 1,626,608 | 1,326,079 | 1,438,102 | 2,367,048 | 2,543,423 | 3,623,777 | 4,139,202 |
| **Total Business** | | **876,234** | **3,254,991** | **2,592,996** | **2,881,707** | **3,146,711** | **2,766,253** | **2,993,949** | **6,805,240** | **7,341,712** |
| Government | Basic research. | 810,749 | 919,965 | 1,961,564 | 1,237,313 | 1,590,686 | N.A. | N.A. | 3,968,097 | 3,893,537 |
| | Applied research. | 1,147,874 | 1,478,384 | 1,145,566 | 2,775,372 | 3,590,181 | N.A. | N.A. | 4,045,337 | 4,058,487 |
| | Experimental development | 1,395,859 | 1,111,992 | 1,061,402 | 1,079,692 | 1,192,893 | N.A. | N.A. | 1,527,797 | 1,515,815 |
| **Total government** | | **3,354,482** | **3,510,342** | **4,168,532** | **5,092,376** | **6,373,760** | N.A. | N.A. | **9,541,230** | **9,467,839** |
| **Higher education** | Basic research. | 1,552,525 | 2,253,839 | 2,262,507 | 1,527,277 | 1,926,086 | 2,701,301 | 2,630,503 | 3,446,437 | 3,903,188 |
| | Applied research. | 2,333,695 | 2,332,512 | 2,100,544 | 1,873,570 | 2,465,347 | 2,673,687 | 2,468,610 | 2,574,179 | 3,007,669 |
| | Experimental development | 652,346 | 1,163,920 | 1,131,495 | 1,469,730 | 1,805,006 | 702,894 | 846,580 | 445,652 | 460,417 |
| **Total higher education** | | **4,538,566** | **5,750,271** | **5,494,547** | **4,870,577** | **6,196,440** | **6,077,882** | **5,945,692** | **6,466,268** | **7,371,274** |
| **Private non profit** | Basic research. | 6,204 | 10,166 | 16,725 | 152,052 | 11,618 | 234,710 | 301,779 | 23,515 | 18,299 |
| | Applied research. | 31,752 | 34,900 | 33,466 | 182,415 | 56,973 | 371,744 | 353,772 | 33,837 | 34,335 |
| | Experimental development | 5,371 | 6,329 | 6,947 | 119,939 | 219,111 | 56,590 | 51,473 | 2,101 | 2,144 |
| **Total private non profit** | | **43,327** | **51,395** | **57,138** | **454,406** | **287,702** | **663,044** | **707,025** | **59,453** | **54,778** |
| **Total** | Basic research. | 2,442,785 | 3,350,719 | 4,413,810 | 3,058,736 | 3,720,873 | N.A. | N.A. | 7,931,309 | 8,372,002 |
| | Applied research. | 3,711,641 | 5,254,887 | 4,072,950 | 6,244,891 | 7,628,629 | N.A. | N.A. | 9,341,555 | 9,746,023 |
| | Experimental development | 2,658,182 | 3,961,393 | 3,826,452 | 3,995,440 | 4,655,112 | N.A. | N.A. | 5,599,327 | 6,117,578 |
| **TOTAL GERD** | | **8,812,608** | **12,566,999** | **12,313,213** | **13,299,067** | **16,004,613** | N.A. | N.A. | **22,872,191** | **24,235,603** |

## Mexican Researcher Fellowships

.

Another important point in order to understand the Mexican Scientific System is the description of the human factor. About twenty years ago, the Mexican government created a Researchers Fellowship (Sistema Nacional de Investigadores-SNI). In this system, the government recognizes the research activity of people in Higher Education, Government Institutions, Private Sector and Non-Profit Organizations. Selection of a fellow is made by a peer review commission. There are two main levels in this fellowship. The lower level is Candidato, addressed to young people starting a researcher career. There are other levels for established researchers (Investigador Nacional). The evolution of the members number of SNI can be seen in Table 4. This table indicates that in Mexico there is less than one researcher per ten thousand habitants (one hundred million is the population in Mexico), with a low rate of young researchers.

TABLE 4

Mexican Researchers Fellowship

| Years | Members | Investigador Nacional | Candidato |
|---|---|---|---|
| 1992 | 6602 | 3947 | 2655 |
| 1993 | 6233 | 3959 | 2274 |
| 1994 | 5879 | 4196 | 1683 |
| 1995 | 5868 | 4309 | 1559 |
| 1996 | 5969 | 4620 | 1349 |
| 1997 | 6278 | 4981 | 1297 |
| 1998 | 6742 | 5513 | 1229 |
| 1999 | 7252 | 5934 | 1318 |
| 2000 | 7466 | 6246 | 1220 |
| 2001 | 8018 | 6890 | 1128 |
| 2002 | 9200 | 7876 | 1324 |

Table 5 shows the distribution of the SNI members according to scientific area.

TABLE 5

Number of members in SNI by scientific area

| Year | Physics, Mathematics and Earth Sciences | Biology and Chemistry | Medicine and Health Sciences | Humanity and Social Sciences | Social Sciences | Biotechnology and Agriculture | Engineering | TOTAL |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | |
| 1992 | 1,099 | 1,363 | 526 | 849 | 575 | 1,218 | 972 | 6,602 |
| 1993 | 1,168 | 1,377 | 527 | 914 | 596 | 836 | 815 | 6,233 |
| 1994 | 1,225 | 1,279 | 563 | 950 | 590 | 572 | 700 | 5,879 |
| 1995 | 1,281 | 1,235 | 586 | 1,022 | 627 | 465 | 652 | 5,686 |
| 1996 | 1,329 | 1,247 | 606 | 1,074 | 663 | 427 | 623 | 5,969 |
| 1997 | 1,436 | 1,314 | 650 | 1,118 | 673 | 463 | 624 | 6,278 |
| 1998 | 1,571 | 1,406 | 703 | 1,172 | 675 | 530 | 685 | 6,742 |
| 1999 | 1,621 | 1,435 | 721 | 1,266 | 738 | 642 | 829 | 7,252 |
| 2000 | 1,569 | 1,435 | 765 | 1,269 | 810 | 700 | 918 | 7,466 |
| 2001 | 1,612 | 1,436 | 846 | 1,362 | 920 | 856 | 986 | 8,018 |
| 2002 | 1,771 | 1,661 | 927 | 1,552 | 1,096 | 1,011 | 1,182 | 9,200 |

Table 6 shows the number of SNI members in some of the main research Institutions.  The second row shows  there is the Total number of fellows in SNI working in State Universities (31) supported by the Mexican Government. The CONACyT research centers row contains all the Research Centers (24) with direct support from CONACyT (Mexican equivalent to NSF).

TABLE 6

Number of members in SNI by Institution

| Institution | Candidato | Investigador nacional | Total | % |
|---|---|---|---|---|
| Universidad Nacional Autónoma de México, UNAM | 189 | 2,385 | 2,574 | 28.0 |
| Public Universities in Different States | 412 | 1,621 | 2,033 | 22.1 |
| CONACyT Research Centers | 139 | 977 | 1,116 | 12.1 |
| Centro de Investigación y Estudios Avanzados, CINVESTAV | 17 | 484 | 501 | 5.4 |
| Universidad Autónoma Metropolitana, UAM | 39 | 499 | 538 | 5.8 |
| National Health Institutes | 77 | 261 | 338 | 3.7 |
| Instituto Politécnico Nacional, IPN | 47 | 256 | 303 | 3.3 |
| Private Universities | 93 | 248 | 341 | 3.7 |
| Instituto Mexicano del Seguro Social, IMSS | 38 | 169 | 207 | 2.3 |
| Colegio de Posgraduados | 26 | 146 | 172 | 1.9 |
| Instituto Nacional de Investigaciones Forestales y Agropecuarias | 12 | 129 | 141 | 1.5 |
| Instituto Nacional de Antropoligía e Historia | 2 | 86 | 88 | 1.0 |
| Others | 233 | 512 | 763 | 8.3 |
| **Total** | 1,324 | 5,385 | 9,200 | 100 |

## 3) OBJECTIVES

Identify the S&T core competencies of Mexico. Further, generate a process that could be used efficiently and rapidly to assess the S&T core competencies in other countries of interest.

## 4) APPROACH AND RESULTS

### 4.1) Overview

Two major types of information are required for a country S&T core competency assessment. One is technical infrastructure, which encompasses the prolific performers, journals that contain many of the papers, the prolific institutions, and the most cited papers/ authors/ journals. The other is technology thrusts, and the relationship among the thrusts. This study focused on obtaining both types of information, using multiple approaches for identifying the thrusts and their relationships. Since the study is a proof-of-principle demonstration, many approaches were examined, and only the most efficient are recommended for future studies. Many labor-intensive manual approaches were used, to serve as benchmarks for validating the more automated approaches. Hopefully, future studies can be performed using the automated or semi-automated approaches.

Human intervention will still be required, but some of the more mechanistic tasks can be handled by computer.

Two types of results are presented, bibliometrics and taxonomies. Bibliometrics provide an indication of the technical infrastructure (prolific authors, journals, institutions, citations), while taxonomies provide an indication of major technology thrusts and their relationships.

Section 4.2 describes the database used for the bibliometrics and taxonomy analyses. Section 4.3 presents the bibliometrics approaches and results, where section 4.3.1 presents the publication bibliometrics, and section 4.3.2 presents the citation bibliometrics. Section 4.4 presents the taxonomy approaches and results, where section 4.4.1 presents the manual taxonomy approaches and results, and section 4.4.2 presents the statistical taxonomy approaches and results.

There are three manual taxonomy approaches and results presented (full Abstract, journal, Keyword phrases), and two major classes of statistical taxonomy approaches and results presented (concept clustering and document clustering). Concept clustering includes factor matrix-based clustering and multi-link hierarchical aggregation clustering. Document clustering includes Greedy String Tiling, entropy-based data compression, partitional, journal, and latent semantic.

### 4.2) Databases and Information Retrieval Approach

For the present study, the Science Citation Index database was used. The retrieved database used for analysis consists of selected journal records (including the fields of authors, titles, journals, author addresses, author keywords, abstract narratives, and references cited for each paper) obtained by searching the Web version of the SCI for articles that contained at least one author with a Mexico address. At the time the final data was extracted for the present paper (Fall 2002), the version of the SCI used accessed about 5600 journals (mainly in physical, engineering, and life sciences basic research).

The SCI database selected represents a fraction of the available Mexican (mainly research) literature, that in turn represents a fraction of the Mexican S&T actually performed. The articles contained within the SCI database do not include the large body of classified literature, or company proprietary technology literature, although the SCI articles could reference these literatures. The SCI articles do not include technical reports, books, or patents from Mexican S&T, but could again reference these literatures. The SCI covers a finite slice of time (2002). The database used represents the bulk of the peer-reviewed high quality Mexican research literature, and is a representative sample of all Mexican research in recent times.

### 4.3) Bibliometrics

A total of 4529 records were retrieved, and the bibliometrics data were extracted.

### 4.3.1 Publication Statistics on Authors, Journals, and Organizations

The first group of metrics presented is counts of papers published by different entities. These metrics can be viewed as output and productivity measures. They are not direct measures of research quality, although there is some threshold quality level inferred, since these papers are published in the (typically) high caliber journals accessed by the SCI.

In all previous text mining studies published by the first author's group, bibliometrics were performed on the overall database retrieved. Since all these previous studies focused on a technology, the resultant bibliometrics provided the technical infrastructure for that technology. In the present case, the focus is on the wide range of technologies being developed within a country. Applying the bibliometrics analysis to the total retrieved database for that country will not provide very useful results. Visitation strategies (one desired application) are typically developed for a specific technology using a group of experts for that technology.

The approach taken here is to identify the thematic thrust areas for the clustering performed in the latter part of this report, then retrieve documents that address each theme. The bibliometrics will then be performed on a theme by theme basis. For the present study, one theme is selected as an illustrative example for the bibliometrics.

Based on the computational linguistics (clustering) results, Thin Films is an important thrust area of Mexican research. A query for thin film research in Mexico was inserted into the Science Citation Index, and 1727 records were recovered for the period 1991-2003, of which 1693 had Abstracts. The bibliometrics analysis was performed on these 1727 records.

#### 3.1.1. Prolific Authors

Table 7 lists the twenty most prolific authors in Mexican thin film research, including their institutions. Two institutions predominate: UNAM and CINVESTAV, IPN. As reference 56 shows, the institution with the most scientists in Mexico is UNAM, followed by CINVESTAV. This institution Center was part of IPN some years ago. The area of Thin Films obeys a similar feature; these two institutions (UNAM and CINVESTAV) do most of the research in this area.

TABLE 7 – MOST PROLIFIC MEXICAN THIN FILM AUTHORS

| AUTHOR NAME | INSTITUTION | #PAPERS |
|---|---|---|
| ZELAYA-ANGEL--O | IPN | 79 |
| NAIR--PK | UNAM | 78 |
| FALCONY--C | CINVESTAV | 71 |
| SEBASTIAN--PJ | UNAM | 70 |
| GONZALEZ-HERNANDEZ--J | CINVESTAV, IPN | 66 |
| NAIR--MTS | UNAM | 55 |
| RAMIREZ-BON--R | UNIV SONORA | 47 |

| | | |
|---|---|---|
| PENA--JL | CTR INVEST CIENTIFICA | 43 |
| ORTIZ--A | UNAM | 42 |
| CASTRO-RODRIGUEZ--R | CINVESTAV, IPN | 40 |
| JERGEL--M | CINVESTAV, IPN | 37 |
| CONTRERAS-PUENTE--G | CINVESTAV, IPN | 37 |
| ASOMOZA--R | CINVESTAV, IPN | 36 |
| JIMENEZ-SANDOVAL--S | CINVESTAV, IPN | 35 |
| ESPINOZA-BELTRAN--FJ | CINVESTAV, IPN | 35 |
| ANDRADE--E | UNAM | 34 |
| ALONSO--JC | UNAM | 32 |
| HARO-PONIATOWSKI--E | UAM-I | 31 |

### 3.1.2. Prolific Journals

Table 8 lists the fifteen most prolific thin film journals containing Mexican research papers. They appear to be top-quality journals, concentrated in physics and materials, with some emphasis on chemistry as well. All but one (Revista Mexicana de Fisica) are English language journals, and Revista Mexicana de Fisica is one of the most relevant peer reviewed physics journals in Latin America (27). It publishes papers in both English and Spanish.

TABLE 8 – MOST PROLIFIC JOURNALS – MEXICAN THIN FILM RESEARCH

| JOURNAL | #PAPERS |
|---|---|
| THIN SOLID FILMS | 147 |
| REVISTA MEXICANA DE FISICA | 80 |
| JOURNAL OF APPLIED PHYSICS | 66 |
| SOLAR ENERGY MATERIALS AND SOLAR CELLS | 62 |
| PHYSICAL REVIEW B | 48 |
| APPLIED SURFACE SCIENCE | 46 |
| APPLIED PHYSICS LETTERS | 36 |
| SEMICONDUCTOR SCIENCE AND TECHNOLOGY | 35 |
| JOURNAL OF VACUUM SCIENCE & TECHNOLOGY A-VACUUM SURFACES AND FILMS | 34 |
| MODERN PHYSICS LETTERS B | 34 |
| SOLID STATE COMMUNICATIONS | 33 |
| JOURNAL OF THE ELECTROCHEMICAL SOCIETY | 32 |
| MATERIALS LETTERS | 27 |
| JOURNAL OF PHYSICS AND CHEMISTRY OF SOLIDS | 25 |
| JOURNAL OF PHYSICS D-APPLIED PHYSICS | 23 |

### 3.1.3. Prolific Institutions and Countries

This section identifies the most prolific institutions producing Mexican-authored thin film papers, and the countries of the most prolific collaborators with Mexican authors of thin film papers.

Table 9A contains a list of the fifteen most prolific institutions.for Mexican-authored thin film papers, and Table 9B contains a list of the eighteen most prolific countries associated with Mexican-authored thin film papers  Two institutions seem to predominate (as found in the case of most prolific authors): UNAM and IPN, as do four countries (USA, Cuba, France, Spain).

As in the case of the affiliation of the most prolific authors, UNAM and CINVESTAV dominate as most prolific institutions, However, some other State Universities (Puebla, Sonora, San Luis Potosi) and some CONACyT Research Centers (CICESE, CIO) seem to have a role in this topic. One important feature of the institutions analysis is that there is no industry involvement. On the other hand, it is noteworthy that non-Mexican Institutions in this table are mainly from developing countries in collaboration with Mexican thin film groups.  This confirms that most of the research on thin solid films in Mexico is dedicated to low cost technology, as it was found in reference 28.

The country collaborations were investigated further.  To ascertain the impact resulting from these collaborations, the citations from different inter-country collaboration sub-sets were determined.  The thin film papers that were published in 1998, and had the following country combinations in their address field (MEXICO-USA; MEXICO-CUBA; MEXICO-FRANCE; MEXICO-SPAIN), were examined for citations.  The average and median citations are listed in Table 9B, in the two right-most columns, next to the respective collaborating countries.

The USA collaborations produced the most citations.  While the median was similar to most of the other countries listed as collaborators, the average was substantially higher. Three of the twenty papers had over twenty cites, while France had only one paper over ten cites, Spain had one paper at ten cites, and Cuba’s best paper had five cites.  While there were modest differences in the citation distributions among the countries, the real difference was the number of highly cited papers.

TABLE 9A–MOST PROLIFIC INSTITUTIONS–MEXICAN THIN FILM RESEARCH

| INSTITUTION | #PAPERS |
|---|---|
| IPN, CINVESTAV | 828 |
| UNIV NACL AUTONOMA MEXICO | 800 |
| UNIV AUTONOMA METROPOLITANA IZTAPALAPA | 110 |
| UNIV AUTONOMA PUEBLA | 102 |
| UNIV LA HABANA | 94 |
| UNIV SONORA | 68 |
| UNIV AUTONOMA SAN LUIS POTOSI | 65 |
| INST NACL INVEST NUCL | 53 |
| CTR INVEST OPT | 45 |
| CNRS | 42 |
| CTR INVEST CIENT & EDUC SUPER ENSENADA | 40 |
| CSIC | 34 |
| INST MEXICANO PETR | 28 |
| CTR INVEST QUIM APLICADA | 24 |

| | |
|---|---|
| BUAP | 23 |

TABLE 9B–MOST PROLIFIC COUNTRIES–MEXICAN THIN FILM RESEARCH

| COUNTRY | #PAPERS | AVER CITES | MED CITES |
|---|---|---|---|
| MEXICO | 1727 | | |
| USA | 246 | 8.3 | 3 |
| CUBA | 103 | 3.2 | 3 |
| FRANCE | 81 | 3.3 | 1 |
| SPAIN | 72 | 5 | 3 |
| ENGLAND | 39 | | |
| UKRAINE | 32 | | |
| GERMANY | 31 | | |
| RUSSIA | 30 | | |
| JAPAN | 29 | | |
| SLOVAKIA | 26 | | |
| INDIA | 20 | | |
| BRAZIL | 18 | | |
| POLAND | 13 | | |
| CANADA | 13 | | |
| ITALY | 12 | | |
| VENEZUELA | 12 | | |
| COLOMBIA | 11 | | |

## 3.2 Citation Statistics on Authors, Papers, and Journals

The second group of metrics presented is counts of citations to papers published by different entities. While citations are ordinarily used as impact or quality metrics (29), much caution needs to be exercised in their frequency count interpretation, since there are numerous reasons why authors cite or do not cite particular papers (30-31).

The citations in all the retrieved SCI papers were aggregated. The authors, specific papers, and journals cited most frequently were identified, and were presented in order of decreasing frequency. It should be emphasized that these citations are from papers in the retrieved database only. Total citations from all papers in the SCI could be substantially larger in some cases.

### 3.2.1. Most Cited First Authors

Table 10 lists the fifteen most cited first authors by Mexican thin film papers, unmodified for self-citations. In contrast to past text mining studies, where there was minimal overlap between most prolific authors and most cited first authors, in the present case, there are seven authors in common between the two lists (NAIR, NAIR, SEBASTIAN, FALCONY, ZELAYA-ANGEL, ORTIZ, RAMIREZ-BON).

While there are a number of factors that could account for the disjointness between the two lists, in other text mining studies the main factor appeared to be that the most prolific authors tended to rarely be first authors (typically ten percent of their SCI papers, or less). Thus, while papers to which they contributed may have received substantial citations, because of their rare appearance as first authors, they did not (on average) accumulate substantial citations as first authors.

An interesting finding occurred in the present case. Consider a highly prolific and most cited author, PK Nair. This author has 94 entries in the SCI. If these entries are listed in order of citations, in descending frequency, then the fractions of papers first authored are as follows (arranged by first ten most cited, second ten most cited, etc): 7/10, 4/10, 3/10, 3/10, 1/10, 2/10, 1/10, 2/10, 1/10, 0/4. Overall, 26% of Nair's papers are first authored, somewhat higher than most prolific authors in other text mining studies. However, a substantial fraction of Nair's most cited papers are first authored, and a much smaller fraction of Nair's least cited papers are first authored. This selectivity is the reason for the high citations.

.

TABLE 10 – MOST CITED FIRST AUTHORS – MEXICAN THIN FILM RESEARCH

| Author | TimesCited |
|---|---:|
| NAIR PK | 231 |
| NAIR MTS | 149 |
| CHOPRA KL | 105 |
| SEBASTIAN PJ | 96 |
| ORTIZ A | 67 |
| ZELAYAANGEL O | 60 |
| ASPNES DE | 59 |
| BRINKER CJ | 58 |
| RAKHSHANI AE | 57 |
| FALCONY C | 55 |
| LUCOVSKY G | 52 |
| BHATTACHARYA RN | 52 |
| RAMIREZBON R | 51 |
| SANCHEZGIL JA | 50 |
| MAJOR S | 47 |

### 3.2.2. Most Cited Papers

Table 11A lists the fifteen most cited papers by Mexican thin film authors (cited by the retrieved papers in the 1727 paper database only). This is a reasonable mix of papers from the past two decades, and reflects a dynamic field of research.

TABLE 11A – MOST CITED PAPERS – MEXICAN THIN FILM RESEARCH

| PAPER | #CITES |
|---|---:|
| BRITT J, 1993, APPL PHYS LETT, V62, P2851 | 44 |

| | |
|---|---|
| KAUR I, 1980, J ELECTROCHEM SOC, V127, P943 | 33 |
| ZELAYAANGEL O, 1994, APPL PHYS LETT, V64, P291 | 31 |
| HODES G, 1987, PHYS REV B, V36, P4215 | 27 |
| CHOPRA KL, 1983, THIN SOLID FILMS, V102, P1 | 26 |
| CHOPRA KL, 1982, PHYS THIN FILMS, V12, P201 | 24 |
| ORTON JW, 1982, J APPL PHYS, V53, P1602 | 21 |
| NAIR PK, 1991, J PHYS D APPL PHYS, V24, P441 | 20 |
| NAIR MTS, 1994, J APPL PHYS, V75, P1557 | 19 |
| ARANOVICH J, 1979, J VAC SCI TECHNOL, V16, P994 | 19 |
| NAIR PK, 1998, SOL ENERG MAT SOL C, V52, P313 | 18 |
| NAIR MTS, 1989, SEMICOND SCI TECH, V4, P191 | 18 |
| DOOLITTLE LR, 1986, NUCL INSTRUM METH B, V15, P227 | 18 |
| MANIFACIER JC, 1976, J PHYS E SCI INSTRUM, V9, P1002 | 18 |
| NAIR PK, 1988, SEMICOND SCI TECH, V3, P134 | 17 |

In this case, two of the most cited papers have been published in one of the best ranked journals in applied physics (APL). Also, it seems to be that Chopra's and Nair’s works play a seminal role in the research of thin films in Mexico.

Table 11B compares the most and least cited Mexican papers published in 1998, retrieved with the same thin film query. There were 157 papers retrieved. Eight papers had twenty or more citations (most cited), 27 papers had zero citations (least cited), and these two groups were compared. The most cited have over 50% more authors and references compared to the least cited. The most cited were all laboratory demonstrations, typically of novel material formation, or material growth and deposition processes. The least cited were mainly laboratory demonstrations, typically of property measurements, but twenty percent were theoretical studies, and ten percent were long-term demonstrations, such as corrosion development. The most cited researchers were all from universities, while the least cited were mainly from universities, but 25% were from research institutes. Interestingly, none of the articles had industry representation. Of the most cited papers’ first authors, five were from Mexico, two were from England, and one from the USA. Other non-first author countries represented were USA (three articles). Of the least cited papers’ first authors, 23 were from Mexico, two from France, one from Wales, and one from Slovakia. Other non-first author countries represented were France (two papers), Slovakia(1), Ukraine(1), Canada(1), Columbia(1), Venezuela(1).

TABLE 11B – MOST AND LEAST CITED PAPERS – 1998

| | AVE | MED | AVE | MED | AVE | MED |
|---|---|---|---|---|---|---|
| | #AUTH | #AUTH | #REF | #REF | #CITES | #CITES |
| MOST CITED | 7 | 6 | 32 | 27.5 | 31 | 27 |
| LEAST CITED | 4 | 4 | 19 | 17 | 0 | 0 |

### 3.2.3. Most Cited Journals

Table 12 lists the fifteen most cited journals by Mexican thin film authors. The highest ranked journals are in applied physics, and others listed focus on materials, with some chemistry. In common with past text mining studies, there is substantial overlap (nine journals) of the most prolific journals with the most cited journals. The list of most cited journals indicates the predominance of four mainly physics journals as the source literature for Mexican research on thin films.

TABLE 12 – MOST CITED JOURNALS – MEXICAN THIN FILM RESEARCH

| JOURNAL | #CITES |
|---|---|
| PHYS REV B | 2072 |
| J APPL PHYS | 1723 |
| APPL PHYS LETT | 1482 |
| THIN SOLID FILMS | 1324 |
| PHYS REV LETT | 952 |
| J ELECTROCHEM SOC | 924 |
| J VAC SCI TECHNOL A | 427 |
| J CHEM PHYS | 376 |
| SOL ENERG MAT SOL C | 368 |
| J CRYST GROWTH | 353 |
| SOLID STATE COMMUN | 338 |
| APPL OPTICS | 324 |
| NUCL INSTRUM METH B | 251 |
| SEMICOND SCI TECH | 248 |
| NATURE | 242 |

### 4.4) Taxonomies

Based on the complete set of 4529 retrieved papers, two types of taxonomies are presented, manual and statistical.  The manual taxonomies require mainly hand-classification of Abstracts, journals, and keywords into categories, whereas the statistical approaches use more computer-based pre-classification.  In both approaches, strong human input is required for final categorization.

#### 4.4.1) Manual

Several record samples were categorized initially in order to obtain rough approximations of the differences between three of the manual techniques.

Table 13 shows the results of the manual categorization of 278 journal article titles.

TABLE 13 – MANUAL CATEGORIZATION OF JOURNAL ARTICLE TITLES

| Manual Characterization of Themes | |
|---|---|
| Physics | 29.9% |
| Biological and Medical Sciences | 33.2% |
| Chemistry | 16.5% |
| Other Topics | 7.1% |
| Agriculture | 4.7% |
| Mathematical and Computer Science | 3.6% |
| Earth Sciences and Oceanography | 2.5% |
| Materials Science | 2.5% |

Table 14 shows the results of the manual categorization of over 400 source journals, covering approximately 3,000 or 60% of the total articles.

TABLE 14 – MANUAL CATEGORIZATION OF JOURNAL TITLES

| Manual Characterization of Journals | |
|---|---|
| Physics | 37.5% |
| Biological and Medical Sciences | 31.0% |
| Chemistry | 11.9% |
| Other Topics | 6.4% |
| Agriculture | 3.6% |
| Mathematical and Computer Science | 3.6% |
| Earth Sciences and Oceanography | 2.6% |
| Material Science | 3.5% |

Table 15 shows the results of the manual characterization of over 300 significant keywords covering approximately 3,000 record counts. The keywords were derived from the identifiers that the article author or editor assigned to the journal articles. Due to the nature of the differences in the editorial policies of the various journals, it can not be stated definitively that all keywords were assigned by the journal article authors.

TABLE 15 – MANUAL CATEGORIZATION OF KEYWORDS

| Manual Characterization of Keywords | |
|---|---|
| Physics | 26.0% |
| Biological and Medical Sciences | 57.6% |
| Chemistry | 10.1% |
| Other Topics | 2.9% |
| Agriculture | 1.8% |
| Mathematical and Computer Science | 0.4% |
| Earth Sciences and Oceanography | 0.6% |
| Material Science | 0.6% |

The topical area percentages in Tables 13 and 14 are roughly similar. Table 15 contains a much higher fraction of Biological and Medical Sciences, due to the larger number of Keywords used in Biological and Medical Sciences papers. This is shown more clearly in the next section.

4.4.1.1) Full Abstract

A random sample consisting of approximately 510 records was extracted from the full database, read individually, and categorized manually. The DTIC taxonomy (See Appendix 1) was used for classification. Approximately 5% of the records were incomplete or otherwise could not be categorized. Table 16 shows the results of this manual effort.

TABLE 16 – MANUAL CATEGORIZATION OF FULL ABSTRACTS

| Manual Characterization of Full Abstract | |
|---|---|
| Physics | 23.1% |
| Biological and Medical Sciences | 34.7% |
| Chemistry | 12.9% |
| Other Topics | 10.5% |
| Agriculture | 4.9% |
| Mathematical and Computer Science | 6.3% |
| Earth Sciences and Oceanography | 5.1% |
| Material Science | 2.4% |

The number of words and identifiers for each Abstract were calculated. Using the source characterization, approximately 4,000 records were grouped to determine the statistics for the Abstracts and identifiers by technical discipline. Table 17 summarizes the Abstract word count statistics:

TABLE 17 – ABSTRACT WORD COUNT STATISTICS

| Article Count | Theme Characterization | Abstract Word Count | | |
|---|---|---|---|---|
| | | Max | Avg | Median |
| 4 | Test Equipment, Research Facilities and Reprograpy | 173 | 114 | 105 |
| 15 | Propulsion, Engines and Fuels | 337 | 152 | 136 |
| 43 | Power Production and Engergy Conversion | 216 | 128 | 129 |
| 845 | Physics | 346 | 116 | 106 |
| 18 | Nuclear Science and Technology | 356 | 166 | 150 |
| 47 | Mechanical, Industrial, Civil and Marine Engineering | 240 | 145 | 145 |
| 204 | Math and Comp Science | 299 | 102 | 91 |
| 145 | Materials | 399 | 144 | 134 |
| 32 | General Science Topics | 275 | 168 | 174 |
| 55 | Environmental Pollution and Control | 457 | 181 | 177 |
| 14 | Electrotechnology and Fluidics | 171 | 98 | 96 |
| 187 | Earth Sciences and Oceanography | 615 | 205 | 189 |
| 407 | Chemistry | 510 | 144 | 140 |
| 8 | Biotechnology | 341 | 206 | 188 |
| 1595 | Biological and Medical Sciences | 672 | 195 | 197 |
| 20 | Atmospheric Sciences | 315 | 174 | 171 |
| 212 | Astronomy and Astrophysics | 513 | 191 | 174 |
| 136 | Agriculture | 472 | 228 | 228 |

Table 18 summarizes the statistics for the identifiers associated with each journal article:

TABLE 18 – IDENTIFIER WORD COUNT STATISTICS

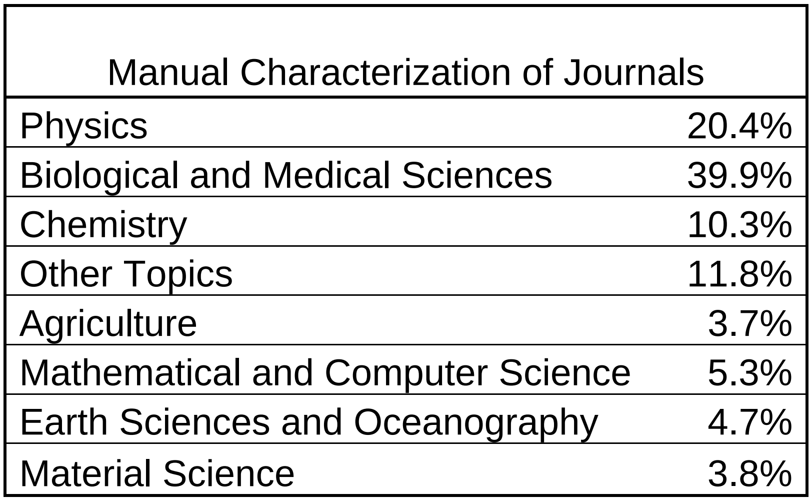
| Article Count | Theme Characterization | ID words | | |
|---|---|---|---|---|
| | | Max | Avg | Median |
| 4 | Test Equipment, Research Facilities and Reprogrampy | 8 | 5.0 | 5.0 |
| 15 | Propulsion, Engines and Fuels | 10 | 4.4 | 3.0 |
| 43 | Power Production and Engergy Conversion | 7 | 2.8 | 2.0 |
| 845 | Physics | 10 | 4.8 | 4.0 |
| 18 | Nuclear Science and Technology | 11 | 3.9 | 2.0 |
| 47 | Mechanical, Industrial, Civil and Marine Engineering | 6 | 2.5 | 2.0 |
| 204 | Math and Comp Science | 10 | 2.8 | 2.0 |
| 145 | Materials | 10 | 3.9 | 3.0 |
| 32 | General Science Topics | 10 | 6.2 | 7.0 |
| 55 | Environmental Pollution and Control | 10 | 5.6 | 6.0 |
| 14 | Electrotechnology and Fluidics | 7 | 3.4 | 3.0 |
| 187 | Earth Sciences and Oceanography | 10 | 5.9 | 5.5 |
| 407 | Chemistry | 13 | 5.7 | 6.0 |
| 8 | Biotechnology | 10 | 7.5 | 8.5 |
| 1595 | Biological and Medical Sciences | 11 | 6.7 | 7.0 |
| 20 | Atmospheric Sciences | 10 | 4.9 | 5.0 |
| 212 | Astronomy and Astrophysics | 10 | 6.9 | 8.0 |
| 136 | Agriculture | 10 | 5.9 | 6.0 |

There are substantial differences in the numbers of Abstract words and keywords among the different disciplines. In particular, Physics, Math, and Computer Sciences articles have about half the number of Abstract words as Biological and Medical Sciences, Agriculture, Earth Sciences and Oceanography articles, and about 70% of the number of Keywords. Correcting Table 15 for this topical cultural difference (some of which is due to many Biomedical journals requiring Structured Abstracts, and their associated additional verbiage) brings the topical ratios in Table 15 more in line with the other manually-derived percentages.

4.4.1.2) Journal

All the journals were manually categorized using the DTIC classification scheme. Those journals that were multi-disciplinary, such as Science or Nature, were classified into a general category. Table 19 shows the results of this manual effort.

TABLE 19 – MANUAL CHARACTERIZATION OF JOURNALS



| Manual Characterization of Journals | |
|---|---|
| Physics | 20.4% |
| Biological and Medical Sciences | 39.9% |
| Chemistry | 10.3% |
| Other Topics | 11.8% |
| Agriculture | 3.7% |
| Mathematical and Computer Science | 5.3% |
| Earth Sciences and Oceanography | 4.7% |
| Material Science | 3.8% |

Table 20 compares the different manual categorization results. If manual categorization of the Full Abstracts is taken as the benchmark, then manual characterization of the Article Titles is the best approximation, and Keyword and Journal Title counts are poorer approximations.

TABLE 20 – COMPARISON OF MANUAL CATEGORIZATION TECHNIQUES

| **Manual Categorization Comparisons** | **Article Titles** | **Journal Titles** | **Keywords** | **Full Abstracts** | **Journals** |
|---|---|---|---|---|---|
| | **Table 7** | **Table 8** | **Table 9** | **Table 10** | **Table 13** |
| Physics | 29.90% | 37.50% | 26.00% | 23.10% | 20.40% |
| Biological and Medical Sciences | 33.20% | 31% | 57.60% | 34.70% | 39.90% |
| Chemistry | 16.50% | 11.90% | 10.10% | 12.90% | 10.30% |
| Other Topics | 7.10% | 6.40% | 2.90% | 10.50% | 11.80% |
| Agriculture | 4.70% | 3.60% | 1.80% | 4.90% | 3.70% |
| Mathematical and Computer Science | 3.60% | 3.60% | 0.40% | 6.30% | 5.30% |
| Earth Sciences and Oceanography | 2.50% | 2.60% | 0.60% | 5.10% | 4.70% |
| Material Science | 2.50% | 3.50% | 0.60% | 2.40% | 3.80% |

The Manual characterization contrasts with the distribution of fellows in SNI by areas, where physics and earth sciences represents less than 20% and Chemistry-Biology and Health sciences are around 30%. This means that the other areas are under represented in terms of Mexican papers appearing in 2002 in the international scientific literature.

### 4.4.2) Statistical Clustering

Two generic types of statistical clustering were used, concept clustering and document clustering. In concept clustering, words or phrases are clustered based on their co-occurrence in the same text unit. In document clustering, documents are clustered based on their overall text similarity.

#### 4.4.2.1) Concept Clustering

Two statistically-based concept clustering methods were used to develop taxonomies, factor matrix clustering and multi-link clustering. Both offer different perspectives on taxonomy category structure from the document clustering approach described later. None of the clustering approaches included here is inherently superior.

In this section, a synergistic combination of factor matrix and multi-link clustering is described that offers substantial improvement in the quality of the resultant clusters. Once the appropriate factor matrix has been generated, the factor matrix can then be used as a filter to identify the significant technical words for further analysis. Specifically, the factor matrix can complement a basic trivial word list (e.g., a list containing words that are trivial in almost all contexts, such as ‘a’, ‘the’, ‘of’, ‘and’, ‘or’, etc) to select context-dependent high technical content words for input to a clustering algorithm. The factor matrix pre-filtering will improve the cohesiveness of clustering by eliminating those words that are trivial words operationally in the application context (32-33).

In addition, the present application compares the use of single words with the use of multi-word phrases for factor generation. There are positives and negatives associated with each approach. Some technical detail is lost by excluding the ordering information contained in multi-word phrases. Conversely, inclusion of all single words compensates for the elimination of some multi-word phrases due to the selection algorithm of the Natural Language Processor. It was desired to examine the trade-off of single words vs. multi-word phrases for factor generation.

### 4.4.2.1.1) Factor Matrix Clustering

#### 4.4.2.1.1.1) Factor Matrix Clustering Approach

Figure 1 is a truncated five factor matrix, shown for illustrative purposes only.

FIGURE 1 – TRUNCATED FIVE FACTOR MATRIX

| Factor | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| receptor | -0.495 | -0.015 | -0.011 | -0.014 | 0.015 |
| receptors | -0.463 | -0.021 | -0.022 | -0.001 | 0.01 |
| antagonist | -0.422 | -0.008 | -0.024 | -0.003 | 0.016 |
| agonist | -0.35 | -0.019 | -0.013 | 0.019 | 0.02 |
| inhibition | -0.347 | -0.018 | -0.016 | 0.017 | 0.02 |
| rat | -0.332 | -0.018 | -0.035 | 0.036 | 0.013 |
| blocked | -0.327 | -0.015 | -0.013 | 0.023 | 0.003 |
| activation | -0.316 | -0.016 | -0.003 | 0.026 | 0.018 |
| binding | -0.313 | -0.01 | -0.007 | 0.004 | 0.018 |
| inhibitor | -0.306 | -0.017 | -0.035 | 0.008 | 0.008 |
| inhibitory | -0.304 | -0.011 | 0.003 | -0.008 | 0.016 |
| inhibited | -0.29 | -0.026 | -0.033 | 0.039 | 0.013 |
| dose-dependent | -0.28 | -0.015 | -0.016 | 0.018 | 0.02 |
| rats | -0.276 | -0.014 | 0.009 | 0.019 | 0.033 |
| stimulation | -0.269 | -0.017 | -0.027 | -0.005 | 0.009 |
| kinase | -0.258 | -0.02 | -0.011 | 0.026 | 0 |
| affinity | -0.246 | -0.023 | 0.004 | 0.016 | 0.017 |
| phosphorylation | -0.24 | -0.021 | -0.004 | 0.035 | -0.004 |
| doses | -0.222 | -0.01 | 0.082 | -0.125 | 0.03 |
| vivo | -0.208 | -0.008 | 0.058 | 0.006 | 0.018 |
| protein | -0.202 | -0.027 | 0.203 | 0.025 | 0.017 |
| toxin | -0.2 | 0.009 | 0.036 | -0.017 | 0.009 |
| brain | -0.197 | -0.001 | -0.035 | -0.055 | 0.007 |
| intracellular | -0.196 | -0.005 | -0.029 | 0.015 | 0 |
| cells | -0.195 | 0.02 | -0.021 | -0.011 | 0.023 |
| Wistar | -0.194 | -0.009 | -0.015 | 0.026 | 0.019 |

The rows contain the words/ phrases and the columns contain the factors. The matrix elements Mij are the factor loadings, or the contribution of word/ phrase i to the theme of factor j. In the example above, the factor loading of the first word (receptor) to the first factor is –0.495. The theme is determined by those words/ phrases that have the largest absolute values of factor loading. Each factor had a positive value tail and negative value tail. For each factor, most of the time, one of the tails dominated in terms of absolute value magnitude. This dominant tail was used to determine the central theme of each factor. In those few cases where the tails were of very similar absolute value magnitude, a theme was extracted from each tail. Thus, Factor 1 in the example above focuses on binding of antagonists to receptors for blocking and inhibition of cell proliferation.

To generate the words/ phrases input to the factor matrix, the highest frequency high technical content words were identified. A factor analysis was performed using the TechOasis statistical package.

After the factor matrices were generated, the word factor matrix was then used for word filtering and selection. In the present study, the words in the factor matrix had to be culled to the approximately 250 allowed by the Excel-based clustering package, WINSTAT. The 250 word limit is an artifact of Excel. Other software packages may allow more or less words to be used for clustering, but all approaches perform culling to reduce dimensionality. The filtering process presented here is applicable to any level of filtered words desired.

The factor loadings in the factor matrix were converted to absolute values. Then, a simple algorithm was used to automatically extract those high factor loading words at the tail of each factor. If word variants were on this list (e.g., singles and plurals), and their factor loadings were reasonably close (32), they were conflated (e.g., 'agent' and 'agents' were conflated into 'agents', and their frequencies were added). A few words were eliminated manually, based on factor loading and estimate of technical content.

4.4.2.1.1.2) Factor Matrix Clustering Results

A list of single words and a list of phrases were generated from the Abstracts using the TechOasis Natural Language Processor. For each list, 1146 high frequency high technical content items were extracted. A factor analysis for words and phrases was performed using the TechOasis statistical package. In each case, a factor matrix consisting of 34 factors resulted. Appendix 2 contains a brief description of each factor in the word factor matrix, and Appendix 3 contains a brief description of each factor in the phrase factor matrix.

The phrases in parentheses represent high factor loading phrases for the factor described, and are presented in inverse order of absolute factor loading value. The decrease in factor loading values is not linear, and the theme of each factor is strongly determined by the first few words/ phrases.

(In the next section, a taxonomy is generated using the multi-link hierarchical clustering approach. The factors in each case above are assigned to the appropriate categories in the taxonomy, providing good coverage and an excellent match.)

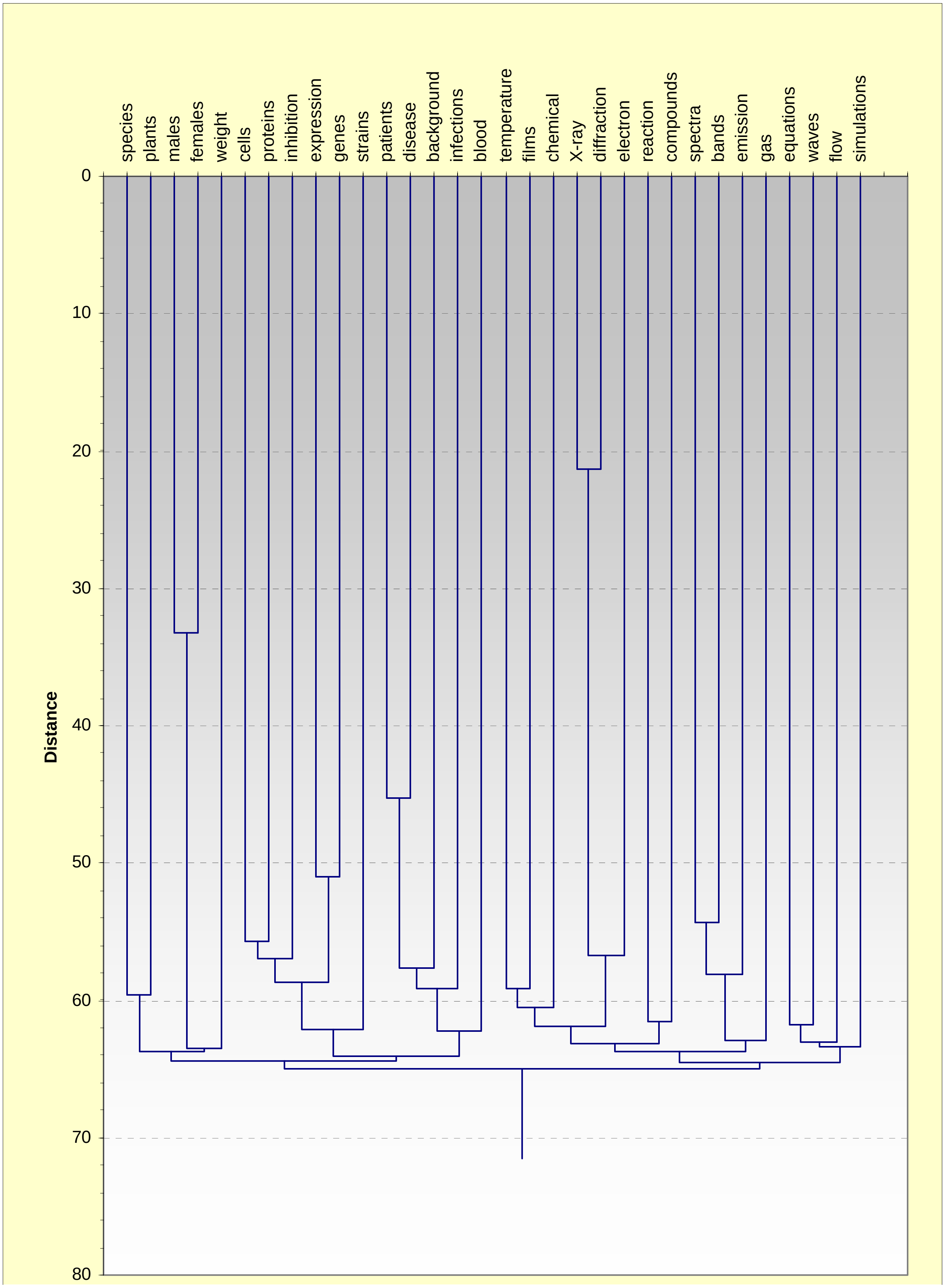
4.4.2.1.2) Multi-Link Hierarchical Word Clustering

4.4.2.1.2.1) Multi-Link Clustering Approach

A symmetrical co-occurrence matrix of the highest frequency high technical content words/ phrases was generated. The matrix elements were normalized using the Equivalence Index (Eij=Cij^2/Ci*Cj, where Ci is the total occurrence frequency of the ith word/ phrase, and Cj is the total occurrence frequency of the jth word/ phrase, for the matrix element ij), and a multi-link clustering analysis was performed using the

WINSTAT statistical package.  The Complete Linkage hierarchical aggregation method was used. A description of the final word dendrogram (a hierarchical tree-like structure), and the aggregation of its branches into a taxonomy of categories, follows in the results section.

Figure 2 is a word-based dendrogram, consisting of 32 high frequency phrases.


species
plants
males
females
weight
cells
proteins
inhibition
expression
genes
strains
patients
disease
background
infections
blood
temperature
films
chemical
X-ray
diffraction
electron
reaction
compounds
spectra
bands
emission
gas
equations
waves
flow
simulations
Distance
0
10
20
30
40
50
60
70
80

One axis is the words, and the other axis ('distance') reflects their similarity. The lower the value of 'distance' at which words, or word groups, are linked together, the closer their relation. As an extreme case of illustration for the dendrogram, words that tend to appear as members of multi-word phrases, such as 'x-ray diffraction', appear adjacent on the dendrogram with very low values of 'distance' at their juncture. In the cluster descriptions that follow, the capitalized phrases in parentheses represent cluster boundary words for each category.

4.4.2.1.2.2) Multi-Link Clustering Results
A description of the final word dendrogram (a hierarchical tree-like structure), and the aggregation of its branches into a taxonomy of categories, follows. See Appendix 4 for the complete word dendrogram.

In the previous focused discipline text mining studies, the Average Link hierarchical aggregation clustering method was used. In those cases, a hierarchical structure could be discerned, and each level of the hierarchy (proceeding downward) described the discipline at increasingly higher levels of detail. In the present country assessment, the clusters are different technologies. A rational hierarchical aggregation at the highest level should not be expected.

In the present study, Complete Link clustering was used rather than Average Link clustering. The top level clusters form a flat set. Some of the clusters have a distinct hierarchical structure into sub-clusters, where a technology area can be divided into its specific sub-technologies. In the cluster descriptions that follow, the capitalized phrases in parentheses represent cluster boundary words for each category. The next section describes the clusters at different levels of the hierarchy, for clusters based on words.

4.4.2.1.2.2.1) Word Clustering Results
The 249 words in the dendrogram are grouped into top level clusters.
At this level, five broad topics can be discerned. These include biology, medicine, physics, chemistry, and environment. Each of these highest level clusters will be divided into smaller clusters, as follows.

1) Biology
There are four main groupings: membrane biology/ cell-cell recognition (CELLS-ESTER); microbial molecular biology/ gene expression (GENES-REGULATORY); recombinant DNA biology (STRAINS-LOCUS); plant population genetics (POPULATIONS-GENOTYPES).

2) Medicine
There are five main groupings: cardiopulmonary (CULTURES-PULMONARY); reproductive (MALES-FOOD); liver damage (BLOOD-FASTING); immunology (INFECTIONS-MARROW); chronic disease treatment (PATIENTS-MELLITUS).

3) Physics

There are four main groupings: quantum and dynamical systems (SIMULATIONS-NEURAL); accelerator physics (SECTIONS-TRANSVERSE); solid-state (SPECTRA-PL); astrophysics (GAS-IONIZED).

4) Chemistry

There are three main groupings: polymers (POWDERS-POLYMERIZATION); molecular characterization (COMPOUNDS-LIGANDS); thin films (TEMPERATURE-BATH).

5) Environment

There are four main groupings: forest and agriculture (SPECIES-SUMMER); oceanography and geophysics (EQUATIONS-DEPTHS); heavy metals in sediments (CU-PB); fish growth (FEED-SHRIMP).

These thematic areas coincide with the major thematic areas listed in Table 20, especially those determined by manual categorization of the full Abstracts.  In Table 20, Agriculture and Earth Sciences and Oceanography were listed as separate themes, whereas the present taxonomy lists them under Environment.

#### 4.4.2.2) Document Clustering

Document clustering is the grouping of similar documents into thematic categories. Different approaches exist (e.g.,34-41). Five approaches were examined in this paper: Greedy String Tiling, Entropy-based Data Compression, Partitional Clustering, Automatic Journal Categorization, and Latent Semantic Clustering.

#### 4.4.2.2.1) Greedy String Tiling

##### 4.4.2.2.1.1 Greedy String Tiling Approach

The approach presented in this section is based on a Greedy String Tiling (GST) text matching algorithm (42-43). It is described in some detail in Appendix 5. Basically, GST clustering forms groups of documents based on the cumulative sum of shared strings of words. Each group is termed a cluster, and the number of records in each cluster, and the highest frequency technical keywords in each cluster, are two outputs central to this analysis.

##### 4.4.2.2.1.2) Greedy String Tiling Results

A five percent similarity threshold produced a total of 1072 clusters. Ninety-three percent of the clusters contained eight Abstracts or less. The 64 largest clusters, (containing 804 Abstracts) were extracted, and are listed in Appendix 6. The main keywords from each cluster (and their frequencies of occurrence within the cluster) are shown in parentheses after the cluster number, and the number of records in each cluster is shown in parenthesis before the cluster number. The keywords are arranged in frequency of appearance, in descending order. Three levels of filtering were used to obtain the main keywords shown below. First, a trivial word list (e.g., of, the, on, etc) was applied to the raw data. Second, only the highest frequency words for each cluster were retained. Third, a manual filtering was performed on the thirty highest words. The themes of each cluster follow the keywords shown.

The taxonomy defined by the word clustering algorithms was used to categorize the 64 clusters generated by the Greedy String Tiling approach. Each cluster was assigned to the most appropriate category in the taxonomy defined by the WINSTAT-generated dendrogram of the last section, based on the theme suggested by the highest frequency technical keywords. The number of records in each taxonomy category from all the clusters in the category was calculated, and is shown in Table 21.

TABLE 21 – ASSIGNMENT OF GST CLUSTERS TO CATEGORIES

| CLUSTER<br>NUMBER | BIOLOGY | MEDICINE | PHYSICS | CHEMISTRY | ENVIRONMENT |
|---|---|---|---|---|---|
| 1 | | | 75 | | |
| 2 | | | | | 26 |
| 3 | | 25 | | | |
| 4 | | | | 19 | |
| 5 | | | | 17 | |
| 6 | | | | 17 | |
| 7 | | | 16 | | |
| 8 | | | 16 | | |
| 9 | | 15 | | | |
| 10 | | 15 | | | |
| 11 | | 15 | | | |
| 12 | | | 13 | | |
| 13 | | 13 | | | |
| 14 | 13 | | | | |
| 15 | | | | | 13 |
| 16 | | | 12 | | |
| 17 | | | | | 12 |
| 18 | | | 12 | | |
| 19 | | | | | 12 |
| 20 | | 12 | | | |
| 21 | | | | 12 | |
| 22 | | 11 | | | |
| 23 | | | | 11 | |
| 24 | | | 11 | | |
| 25 | | | 11 | | |
| 26 | | | 11 | | |
| 27 | | 11 | | | |
| 28 | | | | 11 | |
| 29 | | | | 11 | |
| 30 | | 11 | | | |
| 31 | | | | | 11 |
| 32 | | | | 11 | |
| 33 | | | 10 | | |
| 34 | | 10 | | | |
| 35 | | | 10 | | |
| 36 | 10 | | | | |
| 37 | | | | 10 | |
| 38 | | | | | 10 |
| 39 | | | | 10 | |
| 40 | | 10 | | | |
| 41 | | | | 10 | |
| 42 | | | 10 | | |
| 43 | | | 10 | | |
| 44 | | | 10 | | |

| | | | | | |
|---|---|---|---|---|---|
| 45 | | 10 | | | |
| 46 | | | | 10 | |
| 47 | | | 10 | | |
| 48 | 9 | | | | |
| 49 | | | 9 | | |
| 50 | | | 9 | | |
| 51 | 9 | | | | |
| 52 | 9 | | | | |
| 53 | 9 | | | | |
| 54 | | | 9 | | |
| 55 | | | | 9 | |
| 56 | | 9 | | | |
| 57 | 9 | | | | |
| 58 | | | | 9 | |
| 59 | | | | 9 | |
| 60 | | | | | 9 |
| 61 | | | | | 9 |
| 62 | | | | 9 | |
| 63 | | | | | 9 |
| 64 | | 9 | | | |
| SUM | 68 | 176 | 264 | 185 | 111 |
| SUM (NORM) | 0.08457711 | 0.21890547 | 0.32835821 | 0.2300995 | 0.1380597 |

Compared to the full Abstracts results of Table 16, the present GST categorization provides reasonable agreement in Biology and Medicine (30 vs 34%), modest agreement in Physics (23 vs 33%), and poor agreement in Chemistry (13 vs 23%).

### 4.4.2.2.2) Data Compression Clustering

#### 4.4.2.2.2.1 Data Compression Clustering Approach

The compression algorithm approach (44) of this section assumes that the entropy of a string can be measured when this string is zipped (compressed). The main idea is that when one compresses two strings sequentially, the compression rate will increase if the second string is similar to the first one, and then the zipped string will have less disorder (entropy) than the previous two strings. The entropy is defined as

A)
Entropy = (Length(zip(A+b))-Length(zip(A)) - Length(zip(b+b))+Length(zip(b)) )/ Length(b).

Where A is the patron text, b is the abstract to be analyzed, and zip indicates the zipped function. The fundamental objective is to automate the classification of records into pre-defined categories, such as the DTIC themes. The complete abstract of each record is then compared against the patron text for each pre-determined DTIC theme, and then each record is assigned to an area that provides the best match.

Nineteen patron texts or lexicons for nineteen DTIC themes are defined. With these nineteen DTIC theme dictionaries, the 4529 abstracts are compressed. Then, using the best compression rate, the corresponding first level categorization theme for each abstract is selected. The form to define the patron text and lexicon is explained in Appendix 9.

Two other variants of the Entropy formula are used:

B)
Entropy = (Length(zipL(A+b))-Length(zipL(A))-Length(zipL(b+b))+Length(zipL(b)) )/Length(b).

where zipL indicates a zipping process with the lexicon as parameter. This variant allows shorter calculation time.

C)
Entropy = (Length(zipL(L+b))-Length(zipL(L))-Length(zipL(b+b))+Length(zipL(b)) )/ Length(b).

where the difference is that the Lexicon has been used as a patron text. The computational time is reduced of the order of 6 to 3 hrs. from the A to C Entropy measurement.

#### 4.4.2.2.2.2 Data Compression Clustering Results

Here, it is important to note that with this method it is possible to analyze all abstracts.

The results for automated classification with relative entropy defined by A), B) C) are given in Tables 22A-C.

TABLE 22A
Automated Classification A Formula

| | |
|---|---|
| Physics | 23% |
| Biological and Medical sciences | 32% |
| Chemistry | 8% |
| Agriculture | 8% |
| Mathematical and Computer sciences | 9% |
| Earth sciences and Oceanography | 8% |
| Material sciences | 12% |

TABLE 22B
Automated Classification B Formula

| | |
|---|---|
| Physics | 16% |
| Biological and Medical sciences | 37% |
| Chemistry | 6% |
| Agriculture | 7% |
| Mathematical and Computer sciences | 11% |
| Earth sciences and Oceanography | 4% |
| Material sciences | 19% |

TABLE 22C
Automated Classification C Formula

| | |
|---|---|
| Physics | 16% |
| Biological and Medical sciences | 38% |
| Chemistry | 6% |
| Agriculture | 7% |
| Mathematical and Computer sciences | 11% |
| Earth sciences and Oceanography | 4% |
| Material sciences | 18% |

Although there are some differences between these approaches and the manual characterization, all these results are statistically equivalent to the manual using the Chi-squared statistical test.

4.4.2.2.3) Partitional Clustering

4.4.2.2.3.1 Partitional Clustering Approach

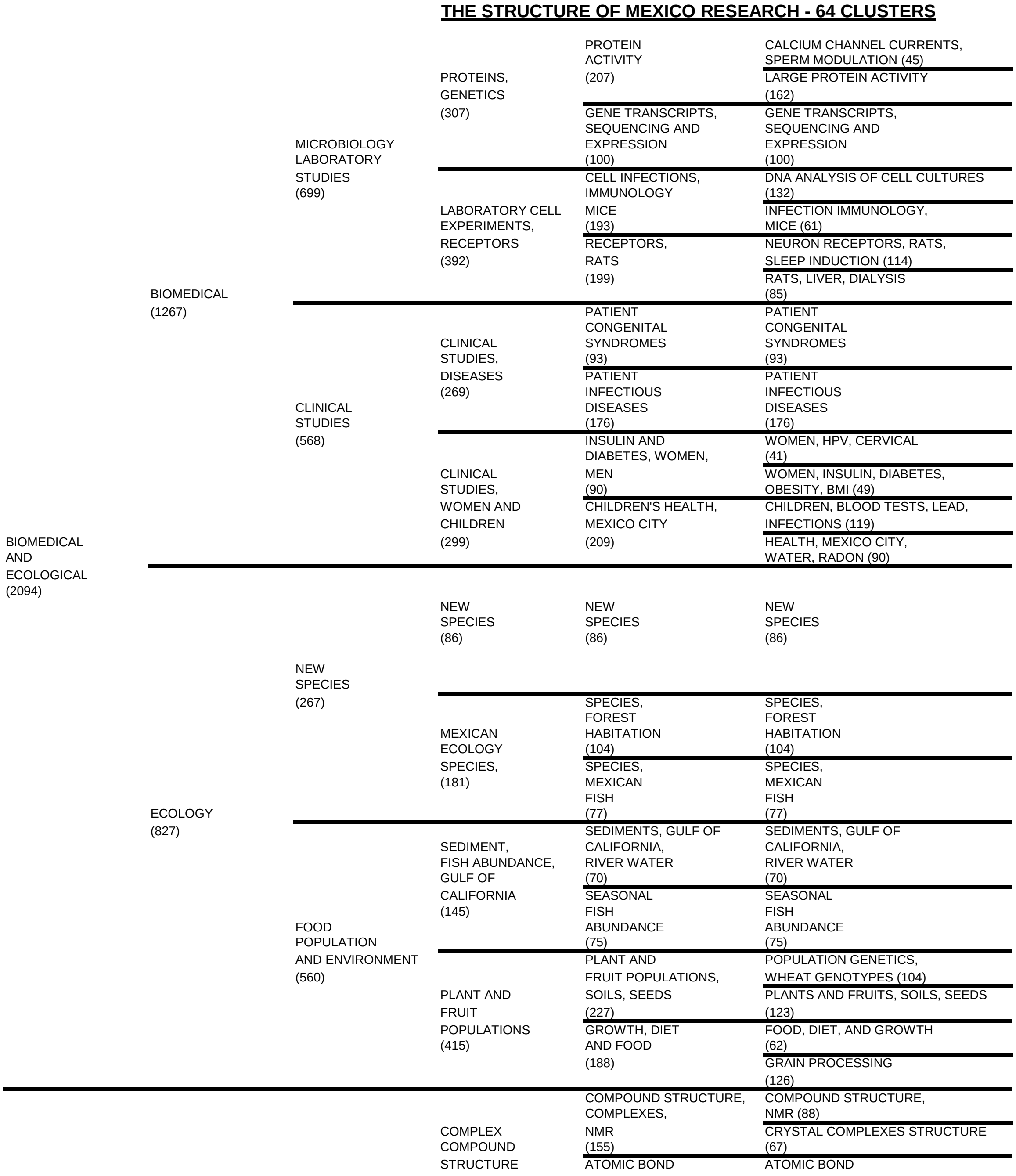
The approach presented in this section is based on a partitional clustering algorithm (53) contained within a software package named CLUTO. Most of CLUTO's clustering algorithms treat the clustering problem as an optimization process that seeks to maximize or minimize a particular clustering criterion function defined either globally or locally over the entire clustering solution space. CLUTO uses a randomized incremental optimization algorithm that is greedy in nature, and has low computational requirements. Appendix 7 describes the partitional clustering approach in more detail.

4.4.2.2.3.2 Partitional Clustering Results

In partitional clustering, the number of clusters desired is input, and all documents in the database are included in those clusters. The 64 clusters that were run are listed in Appendix 8. The main keywords from each cluster (and the percentage of the cluster theme for which they account) are shown in parentheses after the cluster number, and the number of records in each cluster is shown in parenthesis before the cluster number. The keywords are arranged in theme contribution, in descending order.

Three levels of filtering were used to obtain the main keywords shown in Appendix 8. First, a trivial word list (e.g., of, the, on, etc) was applied to the raw data. Second, only the highest frequency words for each cluster were retained. Third, a manual filtering was performed on the thirty highest words. The themes of each cluster (in brief narrative form) follow the keywords shown. The 64 clusters were aggregated into a hierarchical taxonomy using a hierarchical tree generated by the CLUTO software. The taxonomy is shown in Figure 3. The categories in the taxonomy levels, and the number of documents in each category, are described as follows.

FIGURE 3 – PARTITIONAL DOCUMENT CLUSTERING TAXONOMY



**THE STRUCTURE OF MEXICO RESEARCH - 64 CLUSTERS**

| | | | | | |
|---|---|---|---|---|---|
| BIOMEDICAL AND ECOLOGICAL (2094) | BIOMEDICAL (1267) | MICROBIOLOGY LABORATORY STUDIES (699) | PROTEINS, GENETICS (307) | PROTEIN ACTIVITY (207) | CALCIUM CHANNEL CURRENTS, SPERM MODULATION (45) |
| | | | | | LARGE PROTEIN ACTIVITY (162) |
| | | | | GENE TRANSCRIPTS, SEQUENCING AND EXPRESSION (100) | GENE TRANSCRIPTS, SEQUENCING AND EXPRESSION (100) |
| | | | LABORATORY CELL EXPERIMENTS, RECEPTORS (392) | CELL INFECTIONS, IMMUNOLOGY MICE (193) | DNA ANALYSIS OF CELL CULTURES (132) |
| | | | | | INFECTION IMMUNOLOGY, MICE (61) |
| | | | | RECEPTORS, RATS (199) | NEURON RECEPTORS, RATS, SLEEP INDUCTION (114) |
| | | | | | RATS, LIVER, DIALYSIS (85) |
| | | CLINICAL STUDIES (568) | CLINICAL STUDIES, DISEASES (269) | PATIENT CONGENITAL SYNDROMES (93) | PATIENT CONGENITAL SYNDROMES (93) |
| | | | | PATIENT INFECTIOUS DISEASES (176) | PATIENT INFECTIOUS DISEASES (176) |
| | | | CLINICAL STUDIES, WOMEN AND CHILDREN (299) | INSULIN AND DIABETES, WOMEN, MEN (90) | WOMEN, HPV, CERVICAL (41) |
| | | | | | WOMEN, INSULIN, DIABETES, OBESITY, BMI (49) |
| | | | | CHILDREN'S HEALTH, MEXICO CITY (209) | CHILDREN, BLOOD TESTS, LEAD, INFECTIONS (119) |
| | | | | | HEALTH, MEXICO CITY, WATER, RADON (90) |
| | ECOLOGY (827) | NEW SPECIES (267) | NEW SPECIES (86) | NEW SPECIES (86) | NEW SPECIES (86) |
| | | | MEXICAN ECOLOGY SPECIES, (181) | SPECIES, FOREST HABITATION (104) | SPECIES, FOREST HABITATION (104) |
| | | | | SPECIES, MEXICAN FISH (77) | SPECIES, MEXICAN FISH (77) |
| | | FOOD POPULATION AND ENVIRONMENT (560) | SEDIMENT, FISH ABUNDANCE, GULF OF CALIFORNIA (145) | SEDIMENTS, GULF OF CALIFORNIA, RIVER WATER (70) | SEDIMENTS, GULF OF CALIFORNIA, RIVER WATER (70) |
| | | | | SEASONAL FISH ABUNDANCE (75) | SEASONAL FISH ABUNDANCE (75) |
| | | | PLANT AND FRUIT POPULATIONS (415) | PLANT AND FRUIT POPULATIONS, SOILS, SEEDS (227) | POPULATION GENETICS, WHEAT GENOTYPES (104) |
| | | | | | PLANTS AND FRUITS, SOILS, SEEDS (123) |
| | | | | GROWTH, DIET AND FOOD (188) | FOOD, DIET, AND GROWTH (62) |
| | | | | | GRAIN PROCESSING (126) |
| | | | COMPLEX COMPOUND STRUCTURE | COMPOUND STRUCTURE, COMPLEXES, NMR (155) | COMPOUND STRUCTURE, NMR (88) |
| | | | | | CRYSTAL COMPLEXES STRUCTURE (67) |
| | | | | ATOMIC BOND | ATOMIC BOND |

On Figure 3, the columns represent the taxonomy levels. There are six levels depicted in this taxonomy. The highest level (two categories) is the first column, and the lowest level shown (approximately 64 levels) is the last column. The numbers in parentheses represent the number of records assigned to the category.

The first level has two categories: Biomedical and Ecological (2094) and Engineering and Physical Science (2435). Percentage-wise, this is a split of 46/54%. In Table 21 (the manual assignment of GST clusters to categories defined by the word clustering approach), combining the Biology, Medicine, and Environment categories is equivalent to the Biomedical and Ecological category in Figure 3, and combining the Physics and Chemistry categories is equivalent to the Engineering and Physical Science category in Figure 3. In Table 21, the category split of 44/56% compares very favorably with the 46/54% split of Figure 3. In Table 20, the category split of 45/ 55% for the manual clustering of the full Abstracts compares favorably as well.

In Figure 3, the second taxonomy level is generated by sub-dividing each first level category by two. Biomedical and Ecological divides into Biomedical (1267) and Ecology (827), while Engineering and Physical Science divides into Materials and Films (893) and Mathematical, Physics, and Astrophysics Modeling (1542).

Again, comparing Figure 3 with Table 21, Biomedical (from Figure 3) is roughly equivalent to the combination of Biology and Medicine (from Table 21), and Ecology (from Figure 3) is roughly equivalent to Environment (from Table 21). The term 'roughly' is used because sometimes allocation to Biology vs Medicine is not overly clear, or assignment to Biology vs Environment is not overly clear. The Biomedical/ Ecology ratio from Figure 3 (1.53) compares only modestly well with the (Biology & Medicine)/Environment ratio from Table 21 (2.2). The definitional uncertainties are reflected in quantitative differences. Inspection of the GST clusters vs their partitional clustering counterparts shows that these quantitative differences represent manual assigment of clusters to categories vs computer assignment of clusters to categories, more than any intrinsic cluster differences.

Further, Materials and Films (from Figure 3) is roughly equal to Chemistry (from Table 21), and Mathematical, Physics, and Astrophysics (from Figure 3) is roughly equal to Physics (from Table 21). The term 'roughly' is used here because sometimes the allocation to Chemistry vs Physics is not overly clear, especially for materials projects, where the physics of materials and the chemistry of materials are sometimes indistinguishable. The (Materials and Films)/ (Mathematical, Physics, and Astrophysics) ratio from Figure 3 (.58) compares reasonably well with the Chemistry/ Physics ratio from Table 21 (.70). Also, the (Materials and Films)/ (Mathematical, Physics and Astrophysics) ratio from Figure 3 (.58) compares well with the (Chemistry and Materials Sciences)/ (Physics and Mathematical and Computer Science) ratio of full Abstracts from Table 20 (.52).

Three final comments about Figure 3. First, using 64 clusters allows a reasonable picture to be drawn about broad areas of research. If detailed program thrusts were desired, however, many more clusters than 64 would be required. The specific number depends on the degree of focus desired.

For example, from Table 3, the recent Mexico S&T expenditures are on the order of $2.5 Billion/yr. If 64 clusters are used to categorize this S&T, then each cluster (on average) would cover about $40 Million/yr of S&T expenditure. This reflects rather broad categories. If, however, 512 clusters were used, then the resolution increases to about $5

Million/yr for the category average.  This level of resolution would cover small groups of projects.

Second, the Physical, Chemical, and Material Sciences topics identified appear to address forefront areas of research, arreas also addressed by other technologically sophisticated countries.  Research conducted by Mexican scientists somewhat distinctive from that of other countries is concentrated in the Ecology and Biomedicine.  In Ecology, fish, animal, and plant species (and other foods) indigenous to Mexico are focal points, as well as geographical and climatic phenomena.  In Biomedicine, the distinctive aspects focus on health problems indigenous to Mexico, related to geography, environemnt, and diet.

Third, based on the Figure 3 taxonomy, are there any large research gaps evident?  Most of the major research areas appear to be represented, but engineering science (other than materials engineering) does not play a prominent role at the upper taxonomy levels.  As a test, a brief comparison of Mexican and USA papers in a couple of engineering topics was made.  The fraction of Mexican papers that contained the word ‘aircraft” was .00025, while the fraction of USA papers that contained the word ‘aircraft’ was .027, or two orders of magnitude difference.  For the term ‘aerodynamic*, the respective fractions were .00037 and .0137, a factor of 37 difference.

However, here it is important to note that in Mexican science the area of materials science is treated as physics or chemistry. Incorporating the information from the table of Number of members in SNI by scientific area, it is clear that engineering science is underrepresented in the Mexican scientific community, at least as represented in the open technical literature.

### 4.4.2.2.4) Journal Clustering

This section utilizes the ISI classification of journals by categories, and papers are associated in accordance with the category in the ISI.

#### 4.4.2.2.4.1 Journal Clustering Approach

This classification is not in agreement with DTIC, and does not obey criteria as DTIC.

#### 4.4.2.2.4.2 Journal Clustering Results

Automated Classification according to ISI

| | | |
|---|---|---|
| Astronomy | 217 | 0.046229 |
| Atmosphere | 45 | 0.009587 |
| Behavior | 96 | 0.020452 |
| Biology | 1825 | 0.388794 |
| Computer | 63 | 0.013421 |
| Chemistry | 464 | 0.09885 |
| Electronics | 117 | 0.024925 |
| Energy | 63 | 0.013421 |
| Engineering | 101 | 0.021517 |
| Environmental | 170 | 0.036216 |
| Geosciences | 105 | 0.022369 |
| Materials | 276 | 0.058798 |
| Mathematics | 157 | 0.033447 |
| Mechanics | 30 | 0.006391 |
| Multidisciplinary | 32 | 0.006817 |
| Ocean | 92 | 0.019599 |
| Physics | 819 | 0.174478 |
| Radiation | 22 | 0.004687 |

However, these results seem to be in agreement with the manual classification according with DTIC, at least in names.

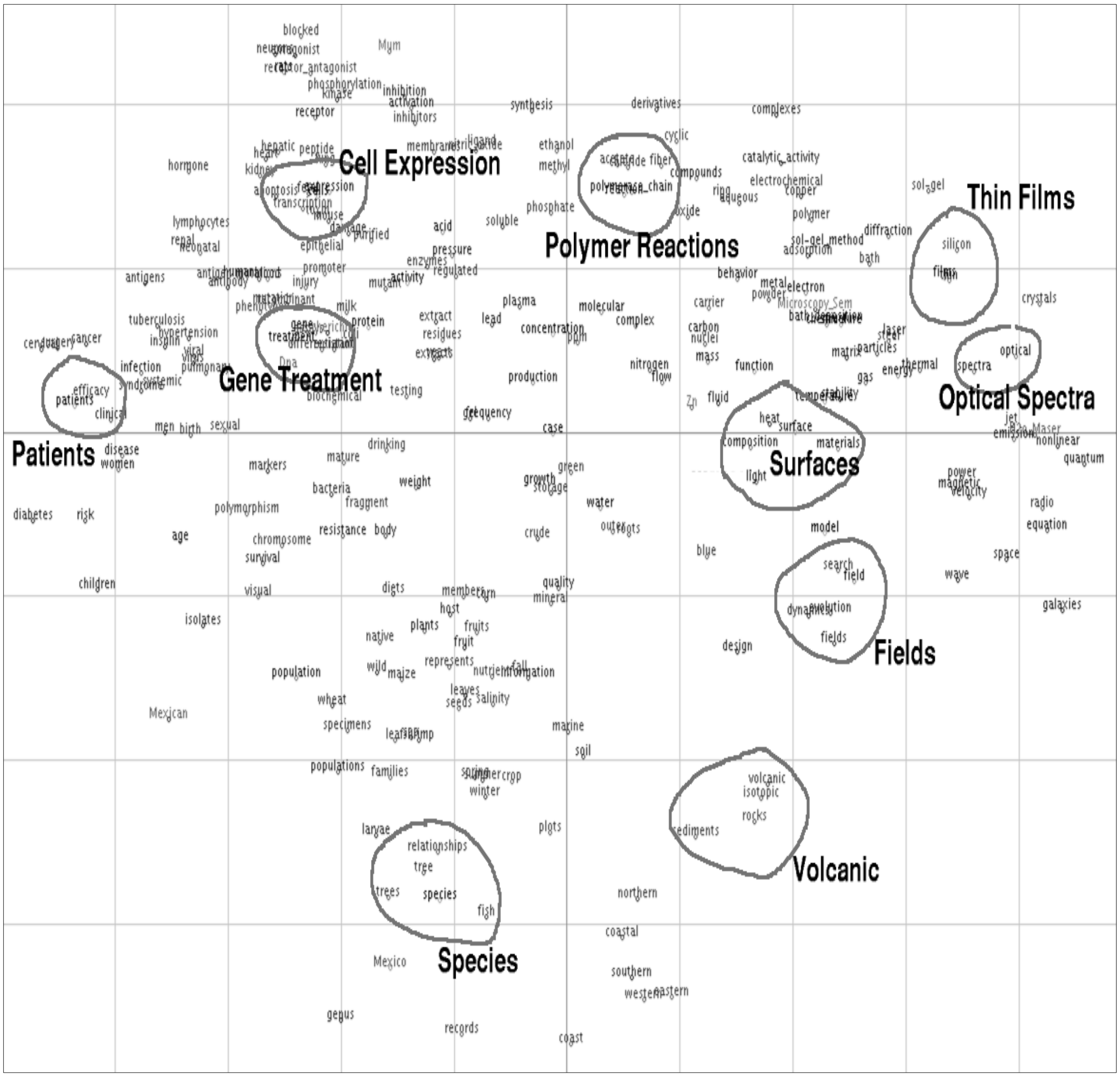
### 4.4.2.2.5) Self-Organising Named Concept Extraction and Clustering

This approach to concept extraction and clustering employs a Bayesian analysis of word co-occurrences, but one which includes nonlinear machine learning algorithms. The method passes through four stages of processing. The first stage involves the seeding of named concepts via extraction of seed terms from the text which possess particular statistical characteristics. The second stage learns a family of related terms around each seeded concept by means of an iterative optimiser with feedback. The result of the first two stages is referred to as a thesaurus, since it bears some resemblance to the thesauri used in Information Science applications. At this stage, the thesaurus has no hierarchy – it is flat. In the third stage, the thesaurus is used to classify the text at a 2-sentence resolution. The tagging of each two sentence segment with multiple concepts generates a directed network of concept co-occurrences. The final stage treats the network of concept co-occurrences as a complex system in order to extract emergent thematic groupings of concepts. This stage results in an interactive visualisation of the concept network. For non-interactive publication, the spatial proximity of clustered concepts and the connectedness of each concept is used to generate a ranked recursive schedule of concept groups. At the lowest level, each concept is described by the lexical term list from the thesaurus.

Below are some examples of thesaurus entries (not in strict rank order), which form the lowest level of the hierarchy:

| ***Concept*** | ***Lexical Terms*** |
|---|---|
| cells | cells Trh internalization Cx43 cell Sertoli transfected macrophage Sf9 lymphocyte germinal dendritic proliferate cancers monocytic |
| species | species helminths Monstrilla subgenus Atlantic_ocean monstrilloid Coreidae Hemiptera tribe synonym Cercidium digenean Qpf niche greggii |
| surface | surface plasmon adsorbed passivation broadening Bet pacificus higher-mode probing Fvc radiometry wafer 4x2 acetylene scribeline |
| films | films thin Cds spray sputtering ellipsometry foils Cdo Cbd Films as-deposited co-sputtering F-7 filamentous Sb2s3-cus |
| acid | acid acetic lactic bell linoleic nucleic uric arachidonic lysophosphatidic demineralization niflumic glutamic aminolevulinic Taurine retinoic |
| gene | gene encodes encoded Streptomyces reporter undetectable exons di-rhamnolipid Drd4 Recr Rhlc St ichthyosis Ais rhamnosyltransferase |
| quantum | quantum dots dilatonic Thomas-fermi exciton undetected excitons reflectometry spins mechanics worlds billiard inter-band polarization-modulation rigorously |

After classification of the data using the thesaurus, and subsequent emergent clustering, a hierarchical concept net was obtained. A screen shot of this, taken from the interactive browser, is shown below:



For the purposes of non-interactive publication, this 2D clustering of the hierarchical network is then serialised into a ranked recursive list of thematic concept groups. Some of these are listed below (not in strict rank order):

| ***Group Name*** | ***Child Groups & Leaf Concepts*** |
|---|---|
| CELLS | cells protein expression treatment gene human blood receptor damage Dna coli Escherichia antibody apoptosis heart recombinant fetal mouse resistant epithelial mutations hepatic mutant milk purified toxin antigen injury promoter biochemical peptide lung assays differentiation phenotype mutation transcription kidney expressing inhibit gland peripheral mitochondrial epithelial_cells regulatory mild actions disorder apoptotic potent saline participation protection organs subunit peripheral_blood initiation pathogenic cells_expressing Western_Blotting |
| SURFACE | surface electron materials chemical bath_deposition composition behavior sol-gel_method gas particles metal matrix laser stability heat Microscopy_Sem adsorption powder polymer bath steel alloy aluminum coatings electrode oxides Sem eta reactor silica reversible Pb Ti ionization chains tau Uv loop microscopic Ftir Cr decomposition surface_tension crude_oil |
| PATIENT | patients disease infection women clinical risk insulin cancer virus men syndrome tuberculosis hypertension cervical antigens birth pulmonary viral surgery efficacy systemic surgical parasite men_women oral care diabetes_mellitus cervical_cancer hospital cardiac birth_weight mycobacterium_tuberculosis systemic_lupus divided_groups multivariate_analysis intestinal_metaplasia pulmonary_tuberculosis patients_underwent |
| OPTICAL | optical emission spectra thermal magnetic H2o_Maser velocity nonlinear power jet radio transverse excited disk Gaas transitions charged tension photon detector formula oscillations mechanics neutron transverse_momentum quantum_wells excited_states phase_transitions porous_media |
| PLANTS | plants body host fruit leaves wild diets corn shrimp maize native spp members salinity seeds fruits leaf represents germination nutrient comparative recovered juvenile nutritional winter_spring white difficult spring_summer segment requirements eggs head crude_protein similarity movement majority superior date white_shrimp |
| SPECIES | species Mexico larvae genus fish tree relationships records trees habitat vegetation seasons genera larval forests |
| SPACE | space galaxies wave radio scalar disk gravity |

| ***Group Name*** | ***Child Groups & Leaf Concepts*** |
|---|---|
|  | compact dual algebra metric formula black_holes matrices expressions scalar_field quantum_wells |

See Appendix 10 for more details.

The interactive version of the full network is currently available from <http://www.leximancer.com/documents/mexico_report/report.html>. Finally, it should be noted that this approach naturally results in classification of the text. This classification system can be used to explore the collection.

4.4.2.2.6) Network Analysis of Word Co-Occurrence

This section presents analysis of Mexico's technology capabilities using network analysis of word occurrence to reveal patterns within the data. These patterns can provide information that would not be evident from an visual examination of the data. This section discusses the data sources and methods, the use of network analysis and the results of the analysis.

**Data sources**

The materials consist of the titles and abstracts of 4,529 documents collected from various sources on the selection criterion of an institutional address in Mexico. Abstracts and titles are studied separately. The research focuses on the abstracts. The abstracts contain 31,724 unique words that occur in total 482,922 times. Titles are used for the comparison (45). The titles contain 10,956 words that occur in total 40,852 times. The title words are used for comparison with the abstract word data.

The title words are packed more densely than the abstract words. Note that the ratio is 40,852/10,956 = 3.73 for title words and 481,922/31,724 = 15.18 for abstract words. This accords with previous research in which one of us has shown that abstract words are less codified than title words (45). Sentences indicating copyright issues were removed from the abstracts. The stopword list available at http://www.uspto.gov/patft/help/stopword.htm was used as a corrective to the inclusion and exclusion of common words. Otherwise, the words were corrected only for the plural 's.'

**Analysis**

An analysis of the data shows that 100 abstract words occur more then 500 times, and that 108 title words occur more then 40 times. In both cases, an asymmetrical matrix was constructed containing the 4,592 documents as the cases and the respective word set as the variables. From this matrix a symmetrical matrix of co-occurrences among the words is generated and a second symmetrical matrix is constructed based on the cosine as a similarity criterion between the words as variables (46-50).[1]

The symmetrical matrices are analyzed using Pajek.[2] The asymmetrical ones are factor analyzed using SPSS.

[1] Salton's cosine is defined as the cosine of the angle enclosed between two vectors *x* and *y* as follows:

$$\text{Cosine}(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{(\sum_{i=1}^{n} x_i^2)*(\sum_{i=1}^{n} y_i^2)}}$$

[2] The homepage of Pajek can be found at http://vlado.fmf.uni-lj.si/pub/networks/pajek/

## Results

### *a. Abstracts*

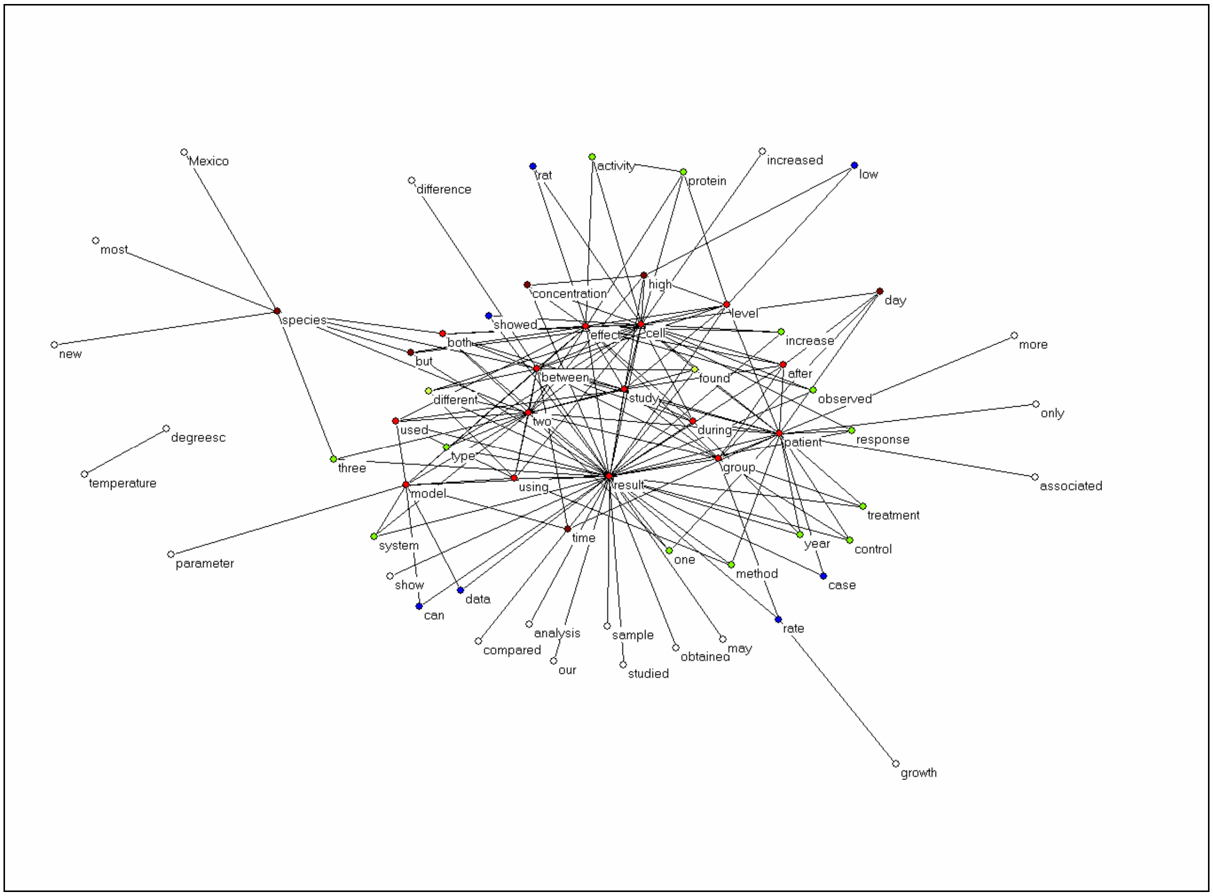


**Figure 4. Co-occurrence map of 63 abstract words co-occurring more than 500 times**.

Sixty-three among the 100 abstract words used co-occur more than 500 times. These are depicted in Figure 4. They form a star shaped network with some interconnecting hubs. The words “effect” and “result” function as hubs and represent the methodologies and their outputs; thus, these results are not highly indicative of capacity. Other words that act as hubs may be more indicative of capacity, including “cell,” “patient,” and “model.” In particular, cell and patient may be aligned with biomedical or biotechnology research.

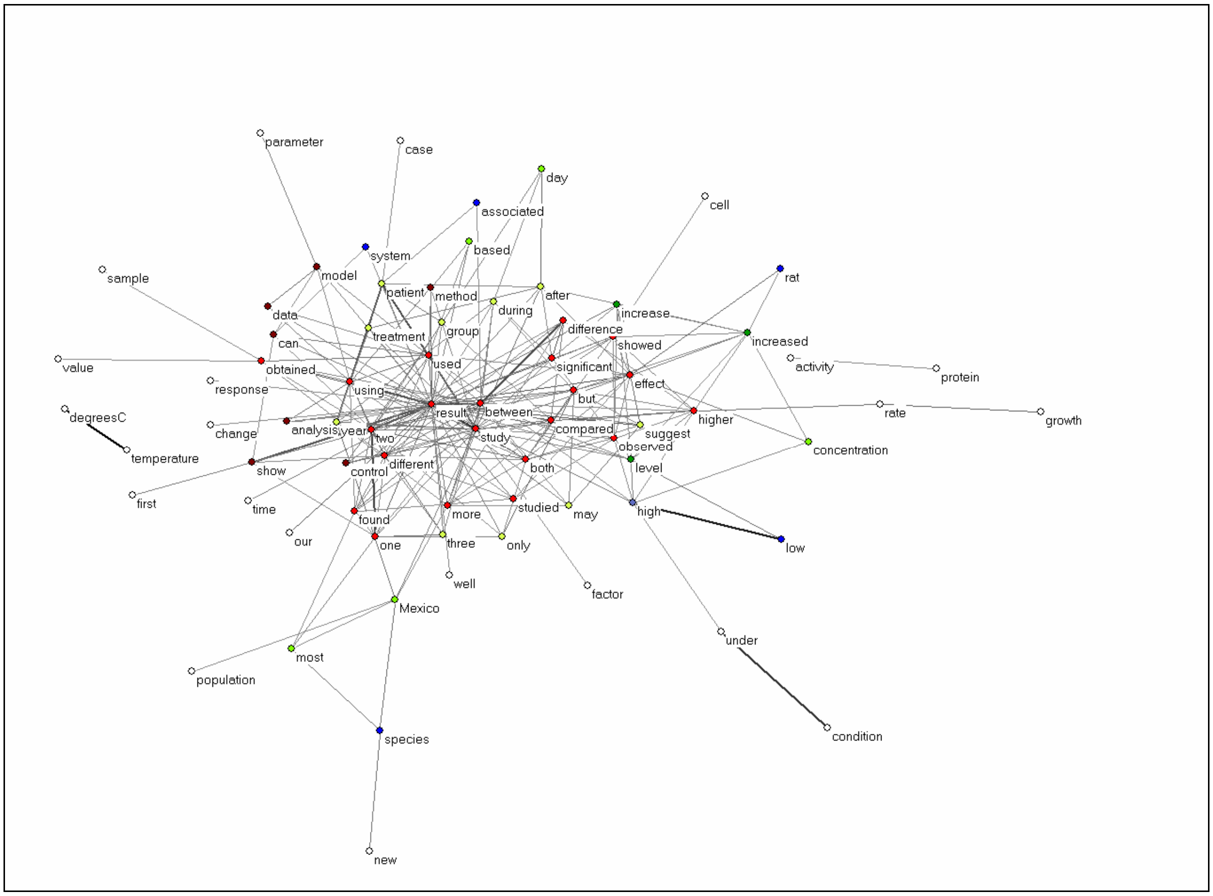


**Figure 5**. Cosine-based map: 74 distributions of abstract words are similar at the level of cosine ≥ 0.2.

Normalization of the word occurrences using the cosine as a similarity criterion (cosine ≥ 0.2) does not change the picture, although some of the stronger relations are highlighted because the star shape is less pronounced. This is depicted in Figure 5.

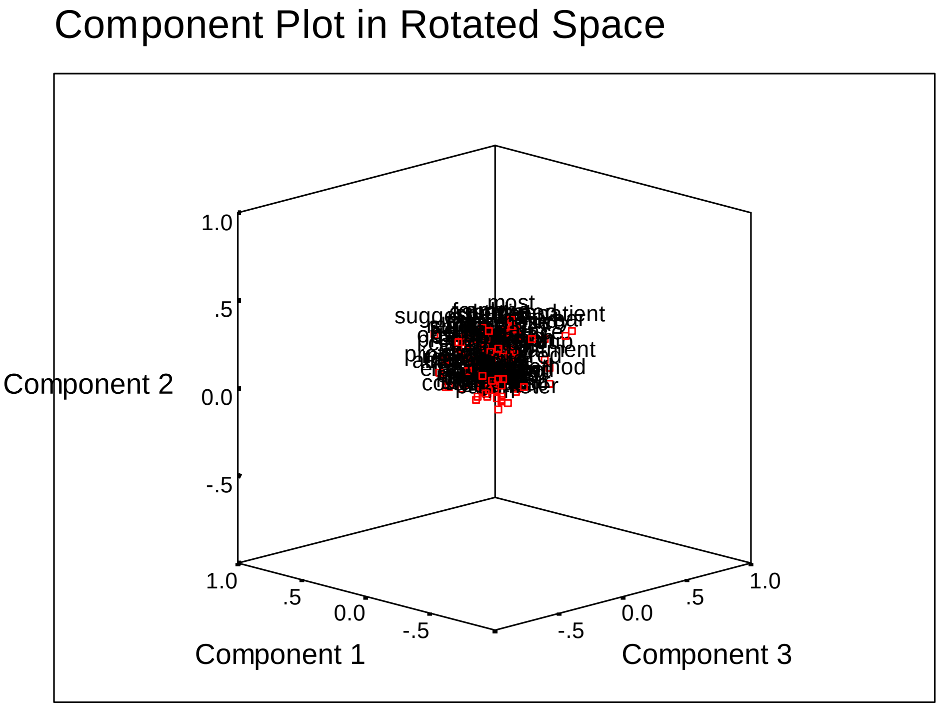


**Figure 6. Plot of loadings on three main factors in a six factor solution using 100 abstract words occurring more than 500 times as variables**

The factor analysis shown in Figure 6 reveals that the normalization does not affect the picture because the matrix is not structured. The number of eigenvalues larger than unity is 36. (These 36 factors explain 50.41 percent of the variance in 100 variables (words).) The redundancy in the distribution of the eigenvalues is only 1.86 percent of the maximum entropy of this distribution (51). This is extremely low.

*b. Title Words*

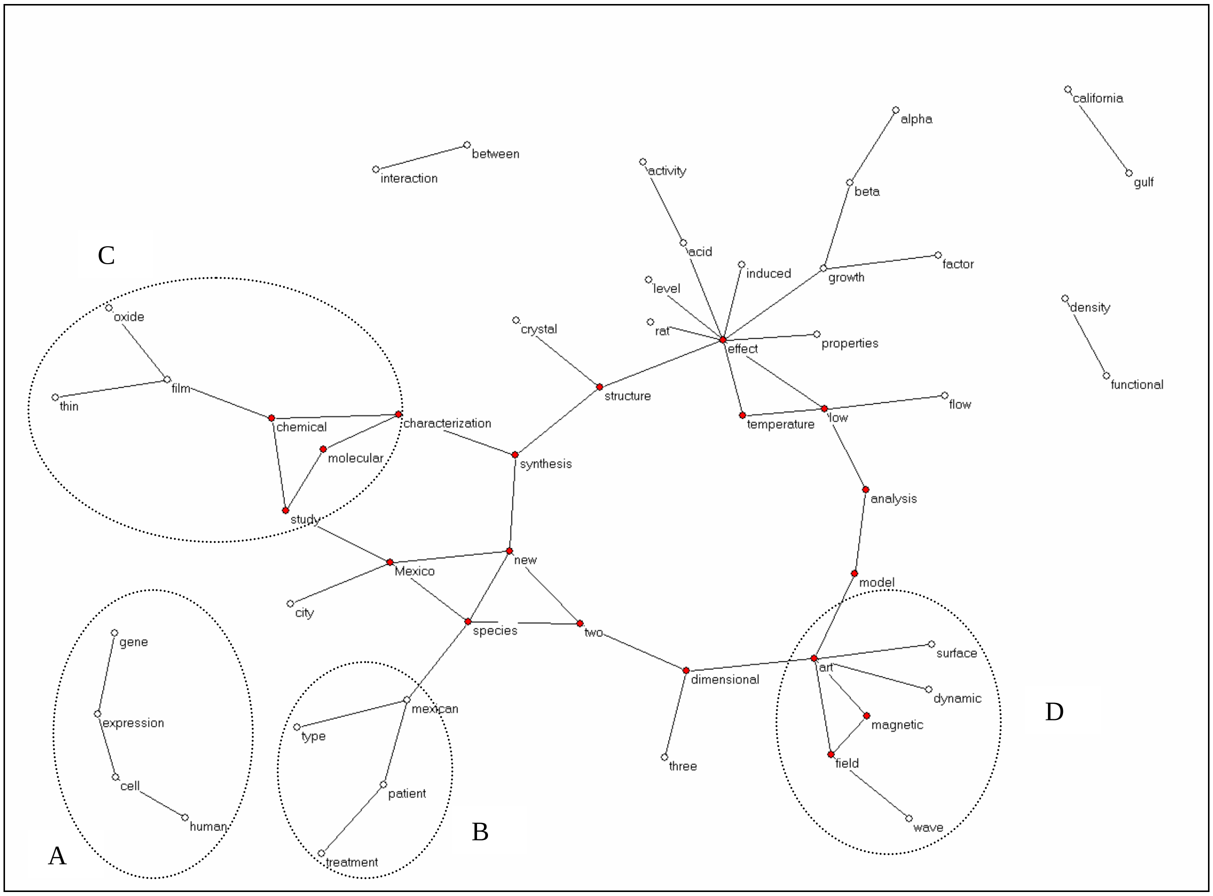


**Figure 7 Co-occurrences with a frequency larger than ten among 53 title words**

Among the 108 title words that occur more than 40 times in the set, 53 words co-occur more than ten times. This network is illustrated in Figure 7. Here, the indication of capacity suggested in the abstract words is upheld by seeing clusters in biotechnology and biomedicine (circles A and B). The biotechnology cluster on the left stands alone (A) disconnected from other title words, although no particular significance can be attached to its position. The medical cluster (B) extending off the bottom of the circle shows Mexican patient treatment as a cluster. The title words also suggest capacity in materials and chemical research -- areas that are not seen to emerge within the abstract words. Note in particular the clusters on the top left (C) which appear to be related to materials, and on the bottom right of the Figure (D) which also appear to be related to chemical or materials research.

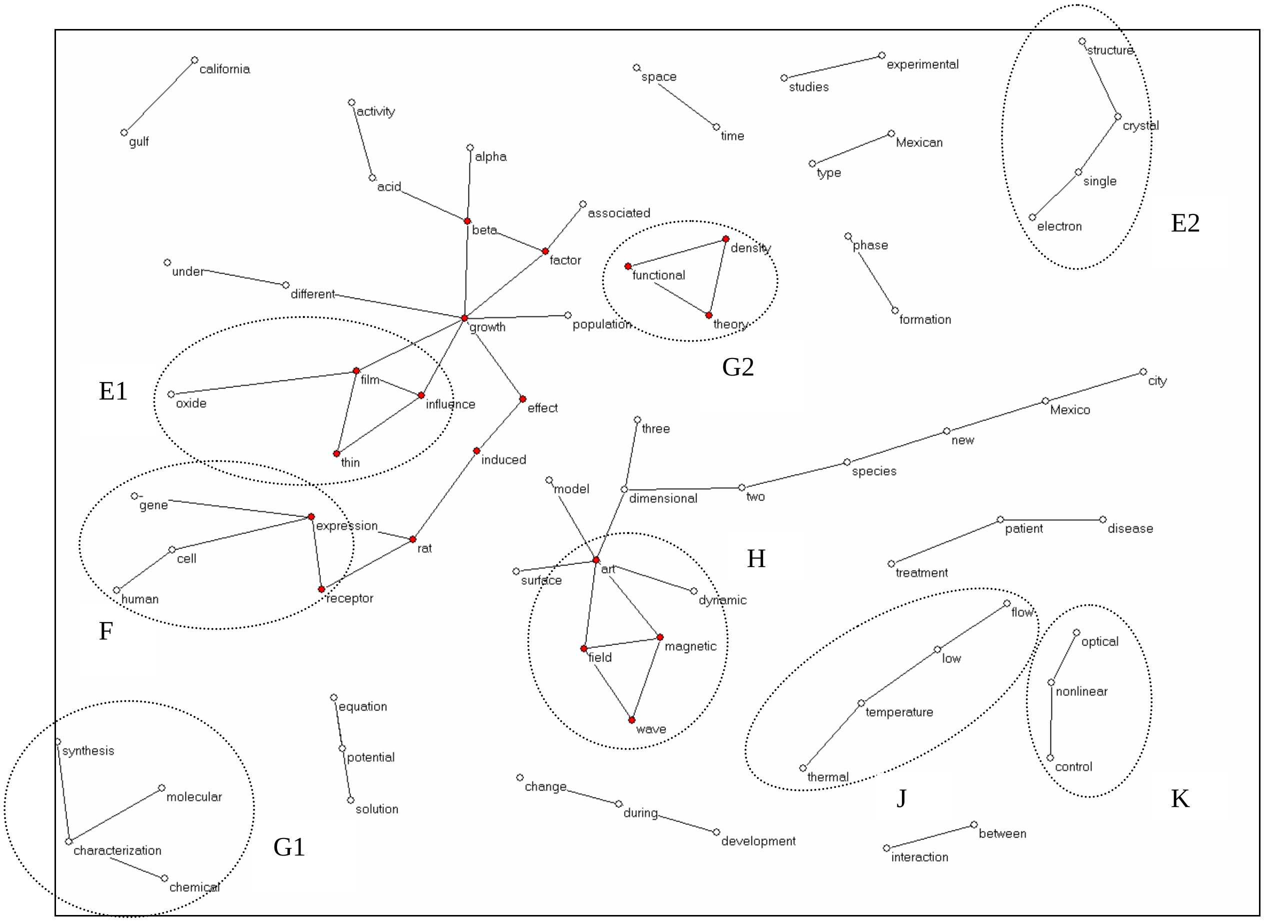


**Figure 8. Cosine-based map: 75 title words are included at the threshold level of cosine ≥ 0.1**

The distribution of the eigenvectors is even flatter than in Figure 8 than is the case with abstract words. The redundancy is smaller than one percent of the maximum entropy (0.84 percent). (48 factors explain 55.28 percent of the variance.) However, the cosine-based map shows that several grouping in the data can be distinguished. This mapping suggests additional capacities over those exposed in Figure 7. The clusters appear to bolster the suggestion drawn from Figure 7 that there are capacities in biomedicine, biotechnology, materials science, and possibly chemistry. These can be further refined to show the possibility of a specialty in materials related to semiconductors (E1 and E2), biotechnology related to genetic expression within human cells (F), and chemical synthesis at the molecular level—nanotechnology?—(G1 and possibly G2).

In addition, this level of analysis suggests several capacities that are not revealed in any other figure. These include a cluster (H) which may suggest capacity in physics and/or astronomy. The cluster revealed in (J) suggests capacities related to semiconductors, polymers and/or geophysics. The cluster (K) also shows a co-occurrence among the words related to optical research, possibly indicating capacities in lasers or other optical research.

## Conclusion

The data is weakly codified. This is a consequence of the selection criterion of the retrieval (i.e., an address in Mexico). Different lines of research are drawn into the set and the result is therefore very heterogeneous. Small groups of co-occurring words can be distinguished in the set of title words, but the abstract words are mainly tied together because of the words related to the word "results."

The structure in the title words can be appreciated as intellectually meaningful despite of the weak structure in the network among the words.  Analysis of the title words are in some ways more suggestive than the abstract words, however, they may be less reliable overall.  Nevertheless, the title words suggest certain capacities within Mexican technology relating to biotechnology, biomedicine, materials research, chemistry, and physics.  This can be checked against overall publications records and citations, which suggest Mexican strength in physics and chemistry (52).

### 4.4.3) Taxonomy Comparisons

Three generic approaches to taxonomy construction were presented: manual clustering, statistical concept clustering, statistical document clustering. The manual clustering of Abstracts was used as the benchmark, and was approximated most closely in the manual group by manual clustering of titles.

The concept clustering approaches (factor matrix, multi-link word/ phrase, self-organizing concept extraction, network analysis) provided complementary perspectives, and all identified the major thrust areas. The document clustering approaches (Greedy String Tiling, Partitional Clustering, Data Compressin, Journal Clustering) showed reasonable agreement among each other, and with the manual Abstract clustering (See table below). The main differences appear to be among Biomed, Chemistry/ Materials, and Environment. Chemical reactions and biological organisms play a role in all three literatures, and slight differences in similarity determination could result in transference of documents among these three clusters.

TECHNICAL CATGORY VS DOCUMENT CLUSTERING TECHNIQUE
(matrix elements in percentages)

| TAXONOMY | **BIOMED** | **PHYSMATH** | **CHEMAT'LS** | **ENVIRONMENT** |
|---|---|---|---|---|
| **GST** | 30.4 | 32.8 | 23 | 13.8 |
| **CLUTO** | 28 | 34 | 19.8 | 18.3 |
| **DATACOMP A** | 32 | 32 | 20 | 18 |
| **DATACOMP B** | 37 | 27 | 25 | 11 |
| **DATACOMP C** | 38 | 27 | 24 | 11 |
| **JOURNALS** | 41 | 34 | 16 | 9 |
| **MANUAL** | 38.6 | 32.7 | 17 | 11.1 |

The first author's very recent unpublished studies on clustering show three main sources of error for all present clustering techniques: 1) excessive trivial words that influence the clustering process; 2) use of different terminology to describe the same concept; and 3) assignment of records to one cluster only. Improved techniques for eliminating trivial words, use of thesauri to normalize terminology, and use of fuzzy clustering to assign individual records to multiple categories would increase the quality of taxonomies substantially. These clustering improvements would also reduce the spread in results of the high quality clustering and network approaches presented in this paper.

## 5. SUMMARY AND CONCLUSIONS

The main objective of this study was to assess the technical core competencies of Mexico. This was accomplished using a variety of clustering approaches. There appear to be four major technical core competencies: Biomedical Sciences includes about 35% of Mexican research; Physics/ Mathematics includes about 30%; Chemistry/ Material Sciences covers about 15%; and Environmental Sciences includes about 10%. The remaining 10% of Mexican research is allocated to myriad other research topics.

The Physical, Chemical, and Material Sciences topics identified appear to address forefront areas of research, arreas also addressed by other technologically sophisticated countries. Research conducted by Mexican scientists somewhat distinctive from that of other countries is concentrated in the Ecology and Biomedicine. In Ecology, fish, animal, and plant species (and other foods) indigenous to Mexico are focal points, as well as geographical and climatic phenomena. In Biomedicine, the distinctive aspects focus on health problems indigenous to Mexico, related to geography, environemnt, and diet.

The Engineering Sciences appear to be under-represented, based on the open source SCI research literature. The Engineering Sciences were not visible at the higher taxonomy levels, and only started to emerge at a few of the lowest level document clusters.

If manual clustering is to be used for taxonomy development, the full Abstract is preferable. If the full Abstract is not available, manual clustering of titles is an acceptable alternative.

The different concept clustering approaches provided complementary perspectives. The factor matrix approach provided good intra-theme word/ phrase quantification linkages, while the network-based approaches provided excellent maps of related concepts.

The document clustering approaches provided good agreement among each other and the benchmark manual Abstract clustering. For more detailed technical analyses, hundreds of clusters would be required. All the document clustering approaches need improvement in handling multi-theme documents and eliminating low technical content words/ phrases. These required improvements are being implemented presently.

The clustering appears useful for generating the structure of a country's S&T, while the bibliometrics appears useful for identifying Centers of Excellence and prolific performers for specific technology areas. Continual upgrades in the clustering algorithms insure that the accuracy of the clusters and categories will continue to improve.

## 6. REFERENCES

1. Kostoff, R. N. Text Mining for Global Technology Watch. In Encyclopedia of Library and Information Science, Second Edition. Drake, M., Ed. Marcel Dekker, Inc. New York, NY. 2003. Vol. 4. 2789-2799.

2. Bostian CW, Brandon WT, Mac Rae AU, Mahle CE, Townes SA . Key technology trends - Satellite systems. Space Communications. 16 (2-3): 97-124 2000.

3. Leneman B. Automation in Soviet Industry, 1970-1983 - An Assessment of the Present State of Robot-Technology. Revue D Etudes Comparatives Est-Ouest. 15 (1): 75-112 1984.

4. Stares P . United-States and Soviet Military Space Programs - A Comparative-Assessment. Daedalus. 114 (2): 127-145 1985.

5. Hutubessy RCW, Hanvoravongchai P, Edejer TTT. Diffusion and utilization of magnetic resonance imaging in Asia. International Journal of Technology Assessment in Health Care. 18 (3): 690-704 SUM 2002.

6. Mooney B, Seymour R . WTEC panels survey Russian maritime technologies. Marine Technology Society Journal. 30 (1): 71-72 SPR 1996.

7. McIntire LV . WTEC panel report on tissue engineering (Reprinted). Tissue Engineering. 9 (1): 3-7 FEB 2003.

8. Robert Campbell, H.D. Balzer, J. Berliner, R. Dobson, and P. Gregory. "Soviet Science and Technology," Foreign Applied Sciences Assessment Center, October 15, 1985.

9. Klinger, A., editor, Klinger, A., et. al., "Soviet Image Pattern Recognition Research," Jan. 1990, Foreign Applied Sciences Assessment Center, *Science Applications International Corp.*, 10260 Campus Point Drive, San Diego, CA 92121, and 1710 Goodridge Drive, McLean VA 22102.

10. *Non-US Data Compression and Coding Research*, R.M. Gray (Ed.), M. Cohn, L.W. Craver, A. Gersho, T. Lookabaugh, F. Pollara, and M. Vetterli, November 1993. A Foreign Applied Sciences Assessment Center (FASAC) report prepared for Science Applications International Corporation (SAIC) under U.S. Government sponsorship.

11. L. J. Lanzerotti, R. C. Henry, H. P. Klein, H. Masursky, G. A. Paulikas, F. L. Scarf, G. A. Soffen, and Y. Terzian, “Soviet Space Science Research,” FASAC Technical Assessment Report FASAC-TAR-3060, Foreign Applied Sciences Assessment Center, 1986.

12. "Soviet Ionospheric Modification Research,” with L.M. Duncan, F.T. Djuth, J.A. Fejer, N.C. Gerson, t. Hagfors, D.B. Newman, Jr., R.L. Showen, Foreign Applied Sciences Assessment Center, Technical Assessment Report 4040, 1988.

13. Spencer, W.J., J.Y. Chen, A. Chiang, W. Frieman, E.S. Kuh, J.L. Moll, R.F. Pease, and K.C. Saraswat, "Chinese Microelectronics," Foreign Applied Sciences Assessment

Center Technical Assessment Report, Science Applications International Corporation, April 1989.

14. R.C. Davidson, M.A. Abdou, L.A. Berry, C.W. Horton, J.F. Lyon, and P.H. Rutherford, Japanese Magnetic Confinement Fusion Research, Foreign Applied Sciences Assessment Center Technical Assessment Report, Science Applications International Corporation, 1990.

15. Kostoff, R. N., "Database Tomography for Technical Intelligence: Comparative Analysis of the Research Impact Assessment Literature and the Journal of the American Chemical Society. Scientometrics, 40:1, 1997.

16. Kostoff, R. N., Eberhart, H. J., and Toothman, D. R. Database Tomography for Technical Intelligence: A Roadmap of the Near-Earth Space Science and Technology Literature. Information Processing and Management. 34:1. 1998.

17. Kostoff, R. N., Eberhart, H. J., and Toothman, D. R. Hypersonic and Supersonic Flow Roadmaps Using Bibliometrics and Database Tomography. Journal of the American Society for Information Science. 50:5. 427-447. 15 April 1999.

18. Kostoff, R. N., Braun, T., Schubert, A., Toothman, D. R., and Humenik, J. Fullerene Roadmaps Using Bibliometrics and Database Tomography. Journal of Chemical Information and Computer Science. 40:1. 19-39. Jan-Feb 2000.

19. Kostoff, R. N., Green, K. A., Toothman, D. R., and Humenik, J. Database Tomography Applied to an Aircraft Science and Technology Investment Strategy. Journal of Aircraft, 37:4. 727-730. July-August 2000.

20. Kostoff, R. N., and DeMarco, R. A. Science and Technology Text Mining. Analytical Chemistry. 73:13. 370-378A. 1 July 2001.

21. Kostoff, R. N., Del Rio, J. A., García, E. O., Ramírez, A. M., and Humenik, J. A. Citation Mining: Integrating Text Mining and Bibliometrics for Research User Profiling. JASIST. 52:13. 1148-1156. 52:13. November 2001.

22. Kostoff, R. N., Tshiteya, R., Pfeil, K. M., and Humenik, J. A. Electrochemical Power Source Roadmaps using Bibliometrics and Database Tomography. Journal of Power Sources. 110:1. 163-176. 2002.

23. Kostoff, R. N., Shlesinger, M., and Malpohl, G. Fractals Roadmaps using Bibliometrics and Database Tomography. Fractals. 12:1. 1-16. March 2004.

24. Kostoff, R. N., Shlesinger, M., and Tshiteya, R. Nonlinear Dynamics Roadmaps using Bibliometrics and Database Tomography. International Journal of Bifurcation and Chaos. 14:1. 61-92. January 2004.

25. 25. Kostoff, R.N., Bedford, C.W., Del Rio, J. A ., Cortes, H., and Karypis, G. Macromolecule Mass Spectrometry: Citation Mining of User Documents. Journal of the American Society for Mass Spectrometry. 15:3. 281-287. March 2004.

26. Kostoff, R. N. Bilateral Asymmetry Prediction. Medical Hypotheses. 61:2. 265-266. August 2003.

27. E. O. García, J. A. del Río, and A.M. Ramírez, Analisis De La Evaluacion De Las Revistas Latinoamericanas A Traves Del Factor De Impacto Renormalizado, Rev. Esp. Doc. Cient. 25, 467-476 (2002).

28. J.A del Río, R.N. Kostoff, E.O. García, A.M. Ramírez y J.A. Humenik Phenomenological Approach To Profile Impact Of Scientific Research: Citation Mining, Adv. Complex Syst. 5, 19-42 (2002).

29. Garfield E. History of citation indexes for chemistry - a brief review. JCICS. 1985; 25(3): 170-174.

30. Kostoff RN. The use and misuse of citation analysis in research evaluation. Scientometrics 1998; 43:1: 27-43.

31. MacRoberts M, MacRoberts B. Problems of citation analysis. Scientometrics 1996; 36(3): 435-444.

32. Kostoff, R. N. The Practice and Malpractice of Stemming. JASIST. 54: 10. June 2003.

33. Kostoff, R. N., and Block, J. A. Factor Matrix Text Filtering and Clustering” JASIST. In Press.

34. Cutting DR, Karger DR, Pedersen JO, Tukey JW. Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR'92). 1992. 318-329.

35. Guha S, Rastogi R, Shim K. CURE: An efficient clustering algorithm for large databases. In *Proceedings of the ACM-SIGMOD 1998 International Conference on Management of Data* (SIGMOD’98). 1998. 73-84.

36. Hearst MA. The use of categories and clusters in information access interfaces. In T. Strzalkowski (ed.), Natural Language Information Retrieval. Kluwer Academic Publishers. 2000.

37. Karypis G, Han EH, Kumar V. Chameleon: A hierarchical clustering algorithm using dynamic modeling. IEEE Computer: Special Issue on Data Analysis and Mining. 1999. 32(8). 68--75.

38. Rasmussen E. Clustering Algorithms. In W. B. Frakes and R. Baeza-Yates (eds.). Information Retrieval Data Structures and Algorithms. 1992. Prentice Hall, N. J.

39. Steinbach M, Karypis G, Kumar V. A comparison of document clustering techniques. Technical Report #00--034. 2000. Department of Computer Science and Engineering. University of Minnesota.

40. Willet P. Recent trends in hierarchical document clustering: A critical review. Information Processing and Management. 1988. 24:577-597.

41. Zamir O, Etzioni O. Web document clustering: A feasibility demonstration. In: Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98). 1998. 46-54.

42. Prechelt L, Malpohl G, Philippsen M. Finding plagiarisms among a set of programs with JPlag. Journal of Universal Computer Science. 2002. 8(11). 1016-1038.

43. Wise MJ. String similarity via greedy string tiling and running Karb-Rabin matching. ftp://ftp.cs.su.oz.au/michaelw/doc/RKR_GST.ps, 1992. Dept. of CS, University of Sidney.

44. Benedetto D, Caglioti E, Loreto V. Language trees and zipping. Physical Review Letters 88 (4): Art. No. 048702 JAN 28 2002.

45. Leydesdorff, L. (1989). Words and Co-Words as Indicators of Intellectual Organization. *Research Policy*, 18, 209-223.

46. Ahlgren, P., B. Jarneving, & R. Rousseau. (2003). Requirement for a Cocitation Similarity Measure, with Special Reference to Pearson's Correlation Coefficient. *Journal of the American Society for Information Science and Technology*, 54 (6), 550-560.

47. Wagner, C. S., & L. Leydesdorff. (2004, forthcoming). Mapping Global Science using International Co-Authorships: A Comparison of 1990 and 2000. *International Journal of Technology and Globalization* (in print).

48. Ortega Priego, J. L. (2003). A Vector Space Model as a Methodological Approach to the Triple Helix Dimensionality: A Comparative Study of Biology and Biomedicine Centres of Two European National Councils from a Webometric View. *Scientometrics,* 58 (2), 429-443.

49. Salton, G., & M. J. McGill. (1983). *Introduction to Modern Information Retrieval*. Auckland, etc.: McGraw-Hill.

50. White, H. D. (2003). Author Cocitation Analysis and Pearson's *r*. *Journal of the American Society for Information Science and Technology*, 54 (13), 1250-1259.

51. Leydesdorff, L. (2003). Meaning and Translation at the Interfaces of Science: Mapping the Case of 'Stem-Cell Research.' Paper presented at the Annual Meeting of the *Society for the Social Studies of Science 4S*, Atlanta, October 2003; at http://www.leydesdorff.net/stemcell

52. Wagner, C.S., & S. Popper (2002). *Technology Use and Productivity in Mexico*, RAND Europe, Final Report.

53. Karypis G.  CLUTO—A clustering toolkit. http://www.cs.umn.edu/˜cluto.

54. Ortuno M, Carpena P, Bernaola-Galvan P, Munoz E, Somoza AM. Keyword detection in natural languages and DNA. Europhysics Letters 57 (5): 759-764 March 2002

55. Reiss H, Hammerich AD, Montroll EW. Thermodynamic Treatment Of Nonphysical Systems - Formalism And An Example (Single-Lane Traffic). Journal Of Statistical Physics 42 (3-4): 647-687 February 1986

**ACKNOWLEDGEMENTS**

The component of work on this paper conducted in Mexico has been partially supported by CONACyT-FOMIX 9250.

# APPENDICES

## APPENDIX 1 – DTIC TAXONOMY

The following contains the top level categories from the DTIC taxonomy. The next level descriptors were used in the categorization, but are too detailed and too lengthy for presentation here.

**CATEGORIES FROM DTIC TAXONOMY**

01--Aviation Technology
02--Agriculture
03--Astronomy and Astrophysics
04--Atmospheric Sciences
05--Behavioral and Social Sciences
06--Biological and Medical Sciences
07--Chemistry
08--Earth Sciences and Oceanography
09--Electrotechnology and Fluidics
10--Power Production and Energy Conversion (Nonpropulsive)
11--Materials
12--Mathematical and Computer Sciences

13--Mechanical, Industrial, Civil and Marine Engineering
14--Test Equipment, Research Facilities and Reprography
15--Military Sciences
16--Guided Missile Technology
17--Navigation, Detection and Countermeasures
18--Nuclear Science and Technology
19--Ordnance
20--Physics
21--Propulsion, Engines and Fuels
22--Space Technology
23--Biotechnology
24--Environmental Pollution and Control
25--Communications

## APPENDIX 2 - WORD FACTOR THEMES

Factor 1 (receptor, receptors, antagonist, agonist, inhibition, rat, blocked, activation, binding, inhibitor, inhibitory, inhibited) focuses on binding of antagonists to receptors for blocking and inhibition of cell proliferation.

Factor 2 (films, thin, substrates, deposition, deposited, film, bath, glass, thickness, coating, coatings, electrical, annealing, annealed, chemical, substrate, gap) focuses on deposition of thin film coatings on substrates, especially using chemical baths.

Factor 3 (diet, diets, dietary, feed, feeding, shrimp, consumption, intake, weight, food, fish, nutritional) focuses on diets for small shrimp, emphasizing impact on feed consumption, weight increase, and nutrition.

Factor 4 (patients, patient, morbidity, mortality, disease, surgery, died, surgical, therapy, age, hospital, background, CI, medical, chemotherapy, cardiac, pulmonary, mellitus, symptoms, diabetes) focuses on hospital studies of patients with advanced chronic diseases, including cancer, cardiovascular diseases, and diabetes.

Factor 5 (star, stellar, galaxy, galaxies, stars, dwarf, NGC, young, radio, emission, Galactic, disk, luminosity, massive, ionized, velocity, telescope, giant, gas, cloud) focuses on observation and mapping of neutral and ionized gas and cloud concentrations and velocities in dwarf galaxies using radio and optical telescopes, emphasizing star formation.

Factor 6 (genetic, polymorphism, gene, allele, DNA, populations, Amplified, cultivars, markers, chromosome, breeding, wheat, locus, diversity, PCR, isolates, hybridization) focuses on allele distributions of DNA polymorphisms in population genetic studies, mainly in humans, but with secondary emphasis on molecular markers of genes with agronomic importance to help breeders facilitate cultivar identification.

Factor 7 (Escherichia, coli, mutant, gene, wild-type, genes, regulatory, promoter, expression, proteins, protein, regulation, encoding, bacteria, sequence, strains, strain) focuses on Escherichia Coli mutant strains.

Factor 8 (Gulf, seismic, basin, zone, fault, km, marine, coastal, coast, vertical, shallow, Sea, depths, ocean, sediments, stations) focuses on the use of seismic techniques for studying crusts and sediments in the water basins surrounding Mexico, and includes the mapping of fault zones.

Factor 9 (nitric, synthase, oxide, ester, methyl) focuses on nitric oxide synthase inhibitors, especially arginine methyl ester.

Factor 10 (rats, Wistar, animals, damage, liver, brain, hepatic, cortex, metabolism, oxidative, rat) focuses on liver and brain damage in experimental Wistar rats and other animals.

Factor 11 (antibodies, infected, infection, ELISA, antigens, antibody, parasite, sera, virus, antigen, viral, immune, infections, assay, assays, human, serum) focuses on detection of immunoglobulin antibody responses to antigens in animals and humans using ELISA.

Factor 12 (forest, forests, species, tropical, tree, trees, habitat, vegetation, habitats, Chiapas, diversity) focuses on diversity of species of trees in tropical forests.

Factor 13 (monomer, polymer, polymerization, pH, polymers, acid, purified, chromatography, chloride, gel, soluble, molecular, chains, sodium) focuses on polymerization of precursor monomers.

Factor 14 (microscopy, diffraction, scanning, electron, X-ray, crystalline, XRD, alloy, powders, SEM, powder, phases, temperature, amorphous) focuses on characterization of crystalline powders and alloys using x-ray diffraction and scanning electron microscopy.

Factor 15 (photoluminescence, emission, spectra, PL, absorption, bands, band, excitation, annealing, nM) focuses on photoluminescence emission and absorption spectra of annealed films, emphasizing band magnitude and spectral shifts.

Factor 16 (NMR, H-1, C-13, X-ray, IR, ligands, spectroscopy, compounds, atoms, atom, complexes, structures, crystal, ligand, bond, compound, diffraction, spectrometry) focuses on characterization of the IR and H-1 and C-13-NMR spectral properties of ligands and compounds, as well as X-Ray crystallographic analyses for the structure of the atoms and complexes.

Factor 17 (insulin, fasting, diabetes, women, cholesterol, glucose, men, serum, Cross-sectional, fat, aged, body, mellitus, intake, blood, plasma, metabolic) focuses on serum insulin, glucose, and cholesterol levels in fasting and non-fasting men and women, and their relation to fat intake and diabetes mellitus.

Factor 18 (plants, seed, seeds, germination, seedling, fruit, leaves, leaf, plant, flowers, fruits, seedlings) focuses on factors that inhibit seed germination in plants, fruit, and flowers.

Factor 19 (Zn, Cu, Ni, Fe, metals, sediments, metal, Mn, Pb, cr, Cd) focuses on concentrations of heavy metals in coastal sediments.

Factor 20 (winter, summer, seasons, spring, season, rainy, dry, annual, biomass, climatic, salinity, nutrient, precipitation, dissolved, ecological) focuses on changes in ecological variables over different climatic seasons.

Factor 21 (cooling, flow, fluid, temperature, heat, gas, heating, thermal, pressure, fluids, volcanic, temperatures, mixing, gases, velocity, equilibrium, flows, thermodynamic) focuses on high temperature and pressure fluid flows, especially of volcanic materials.

Factor 22 (catalytic, catalyst, catalysts, FTIR, Fourier, transform, sulfur, oxides, SEM, reaction, pore, XRD, infrared, reactions) focuses on characterization of catalysts by FTIR, SEM, and XRD, including incorporation of sulfur oxides into the catalyst.

Factor 23 (soil, crop, Zea, soils, erosion, land, agricultural, plots, maize, productivity, biomass, wheat, microbial, irrigation, conservation) focuses on erosion of soil by corn crops, and methods to reverse effects of erosion and improve agricultural productivity by land plot experiments.

Factor 24 (cross, section, collisions, transverse, sections, detector, momentum, bar, jet, Scattering, angular, photon, energies) focuses on inclusive jet collision cross sections, emphasizing transverse momentum and energy dependence.

Factor 25 (females, males, female, sexual, male, age, women, reproductive, sex, reproduction, birth, adult, pregnancy, genital, maternal) focuses on sexual and age factors that influence reproduction.

Factor 26 (polymerase, chain, mutations, reaction, PCR, gene, mutation, amplification, RNA, mRNA, transcription, expression, liver, patients) focuses on use of polymerase chain reaction to analyze gene mutations.

Factor 27 (networks, neural, Monte, Carlo, network, simulation, algorithm, simulations, controller, computer, feedback, noise, robust, image, model, algorithms, nonlinear, signals, control) focuses on Monte Carlo simulations of neural networks for feedback control.

Factor 28 (laser, wave, pulses, field, waves, polarization, beam, pulse, optical, amplitude, magnetic, wavelength, electric, angle, electromagnetic, electron, light, radiation) focuses on pulse generation in laser systems.

Factor 29 (culture, strains, bacteria, bacterial, fermentation, cultures, l(-1, microbial, pH, enzyme, glucose, isolates, strain, electrophoresis, growth) focuses on analysis of bacterial strains from cultures.

Factor 30 (Ca2, channels, currents, Channel, membrane, oocytes, Na, muscle, intracellular) focuses on currents in calcium channels in muscle membranes, emphasizing intracellular calcium release.

Factor 31 (atomic, charge, electronic, density, electron, hydrogen, energy, Carlo, Monte, atoms, energies, bond, electrode, GaAs, transport, molecules) focuses on atomic charges and electronic charge densities.

Factor 32 (apoptosis, cell, apoptotic, proliferation, cells, lymphocytes, vivo, death, marrow, vitro) focuses on lymphocytic cell proliferative and apoptotic processes.

Factor 33 (epithelial, Carlo, capillary, Monte, lymphocytes, inflammatory, vascular, immune, blood, right, animals, smooth, left, porous, fluids) focuses on lymphocyte and epithelial cells in the immune reaction to inflammation, especially in the smooth vascular system.

Factor 34 (quantum, equation, dynamical, Hamiltonian, equations, harmonic, motion, scalar, wave, space, gravitational, gravity, symmetry, cosmological, amplitude, dynamics, field) focuses on modeling of dynamical motions of quantum systems.

## APPENDIX 3 – PHRASE FACTOR THEMES

Factor 1 (peripheral blood, cells, lymphocytes, IL-4, DNA damage, bone marrow, immune system, immune response, proliferation, mice, infection, T cells) focuses on characterization of immune system and immune response by lymphocyte and interleuken concentrations in peripheral blood cells.

Factor 2 (films, glass substrates, thin films, bath, CuS, chemical bath, resistivity) focuses on thin films generated by chemical baths on glass substrates.

Factor 3 (apoptosis, cell death, cell proliferation, p53, proliferation, flow cytometry, tumors, cancer) focuses on cell proliferative and apoptopic (programmed cell death) processes, especially related to cancer.

Factor 4 (Cu, Zn, Ni, Fe, Cd, PB, metals, Cr, CO, sediments, Mn, AG, heavy metals, V) focuses on heavy metals in crustal sediments.

Factor 5 (females, males, sexes, male, species, Yucatan, dry season, Campeche, feeding, sex, reproduction) focuses on reproductive and feeding habits of species of both sexes as a function of location and season.

Factor 6 (striatum, oxidative stress, hippocampus, cerebellum, free radicals, lipid peroxidation, neurons, rats, dose-dependent manner, brain) focuses on study of oxidative stress from free radicals, especially lipid peroxidation, emphasizing cytology in the hippocampus and cerebellum.

Factor 7 (BMI, body mass index BMI, obesity, women, insulin resistance, men, blood pressure, diabetes, insulin, triglycerides, hypertension, baseline, P, cholesterol, age, type 2 diabetes) focuses on relation of Body Mass Index, obesity, insulin resistance, blood pressure, triglycerides, and cholesterol to diabetes, especially Type 2.

Factor 8 (population growth, mL(-1, Chlorella, rotifers, algae, controls, food, yeast) focuses on population growth studies of rotifers for different green algae Chlorella concentrations.

Factor 9 (SLE, IgG, SLE patients, systemic lupus erythematosus SLE, affinity chromatography, IgA, serum samples, ELISA, sera, serum) focuses on immunoglobulin concentrations in the serum samples of systemic lupus erythematosus patients.

Factor 10 (manganese, iron, zinc, copper, chromium, selenium, vitamins, cobalt, cadmium, aluminum) focuses on vitamins and minerals in the diet.

Factor 11 (dwarf galaxies, star formation, stars, dark, H II regions, galaxy, stellar populations, masses, H, mass, galaxies, gas) focuses on star formation and stellar populations in dwarf galaxies.

Factor 12 (legumes, erosion, corn, soil, treatments, DC, plots, farmers, organic, nematodes, insects, soil erosion, diversity, tortillas, germination) focuses on prevention of corn-induced soil erosion by using legumes as soil cover, and studying toxic effects of legumes on insects and nematodes.

Factor 13 (C-13, H-1, IR, solid state, Te, N-15, compounds, mass spectrometry, IR spectra, complexes, selenium, DMSO, ligands, sulfur, crystal, X-ray crystallography) focuses on characterization of the IR and H-1 and C-13-NMR spectral properties of ligands and compounds, as well as X-Ray crystallographic analyses for the structure of the atoms and complexes.

Factor 14 (lectins, adhesion, pathogen, binding, molecular mechanisms, pathogens, fungus, parasite, fungi, recognition, modulation, uptake) focuses on use of lectins to study pathogen adhesion and binding mechanisms.

Factor 15 (diet, shrimp, juveniles, diets, growth rate, survival, growth, salinity, consumption, carbohydrates, digestibility, nutrients) focuses on diets for small shrimp, emphasizing impact on feed consumption, weight increase, and nutrition.

Factor 16 (rats, administration, inhibition, stimulation, GABA, regulation, activation, control rats, rat, inhibitor, expression) focuses on administration of GABA to stimulate and inhibit hormone secretion.

Factor 17 (Mg2, K, Na, Ca2, Zn2, mitochondria, Western blot analysis, size distribution, cytoplasm, Mg, respiration) focuses on cations, especially as enzyme activity inhibitors.

Factor 18 (leaves, flowers, plants, seeds, roots, stems, fruit, fruits, progeny, stem, greenhouse conditions, plant, trees, flowering, germination, seedlings, ammonium, shoots) focuses on germination and growth of plants, flowers, and fruits.

Factor 19 (infection, children, mothers, cervical cancer, gestational age, newborns, women, risk factors, age, HPV infection, ELISA, vaccination, background, pregnancy, antibodies, infants, serum samples, weight, birth) focuses on relation of infection to cervical cancer, especially HPV infection, and development of vaccines to protect against cervical cancer.

Factor 20 (furnace, bath, steel, slag, mathematical model, injection, gases, iron) focuses on mathematical models of reduced iron heating in electric arc furnaces for steel-slag systems.

Factor 21 (crystalline structure, acid, infrared spectroscopy, electron microscopy, texture, x-Ray powder diffraction, X-ray diffraction, crystallite size, compression, electron microscopy SEM) focuses on characterization of crystalline powders and alloys using infrared spectroscopy, x-ray diffraction, and scanning electron microscopy.

Factor 22 (mortality, patients, morbidity, complications, surgery, background, chemotherapy, age, disease, infections, odds ratio, death, multivariate analysis, radiotherapy) focuses on hospital studies of patients with advanced chronic diseases, mainly cancer, emphasizing treatments such as surgery, chemotherapy, and radiotherapy.

Factor 23 (annealing temperature, annealing, optical absorption spectra, composite films, films, different temperatures, optical properties, nanoparticles, optical absorption, ZnO, Rutherford, red, glass) focuses on optical absorption of thin films annealed at different temperatures, especially optical absorption spectra as a function of imbedded nanoparticle size and clusters.

Factor 24 (volcano, Popocatepetl volcano, SO2, eruption, emissions, convection, crystallization) focuses on emissions from erupted Popocatepetl volcano.

Factor 25 (benzene, toluene, urea, aqueous solutions, higher temperatures, solubility, water, cavities, catalyst, entropy, structural changes, temperature, hydrocarbons) focuses on thermodynamic behavior of non- polar solutes in water and in aqueous solutions of protein denaturants.

Factor 26 (Q(2, HERA, cross section, jets, cross sections, jet, Fermilab Tevatron Collider, Pt, photons, bar collisions, x, W, elastic, proton, detector) focuses on inclusive jet collision cross sections, emphasizing high energy scattering at HERA and the Fermilab Tevatron Collider.

Factor 27 (SEM, XRD, TEM, N-2, precursors, coatings, catalysts, deposits, nitrogen, hydrogen, catalytic activity, sulfur, crystal structure, X-ray diffraction XRD) focuses on characterization of properties and structure of coatings using SEM, XRD, and TEM.

Factor 28 (oysters, Bay, Gulf, mouth, lipids, islands, coast, muscle, carbohydrates) focuses on study of oyster cultures and growth in various bodies of water.

Factor 29 (mutant, gene, Mycobacterium tuberculosis, enzymes, genes, expression, Escherichia coli, proteins, PCR, mutations, sequence analysis, tuberculosis, bacteria, strains, toxin, biosynthesis) focuses on gene regulation in mycobacteria.

Factor 30 (acetaldehyde, formaldehyde, hepatocytes, IL-6, ethanol, acetone) focuses on inflammatory mediators for chronic lever disease.

Factor 31 (Br, HF, red, Se, Sr, algae, Baja California Sur, ba, density functional theory, electronic properties, u, tissue, Cr, Ni, AG) focuses on concentration of elements in seaweed, emphasizing density differences in red, brown, and green algae.

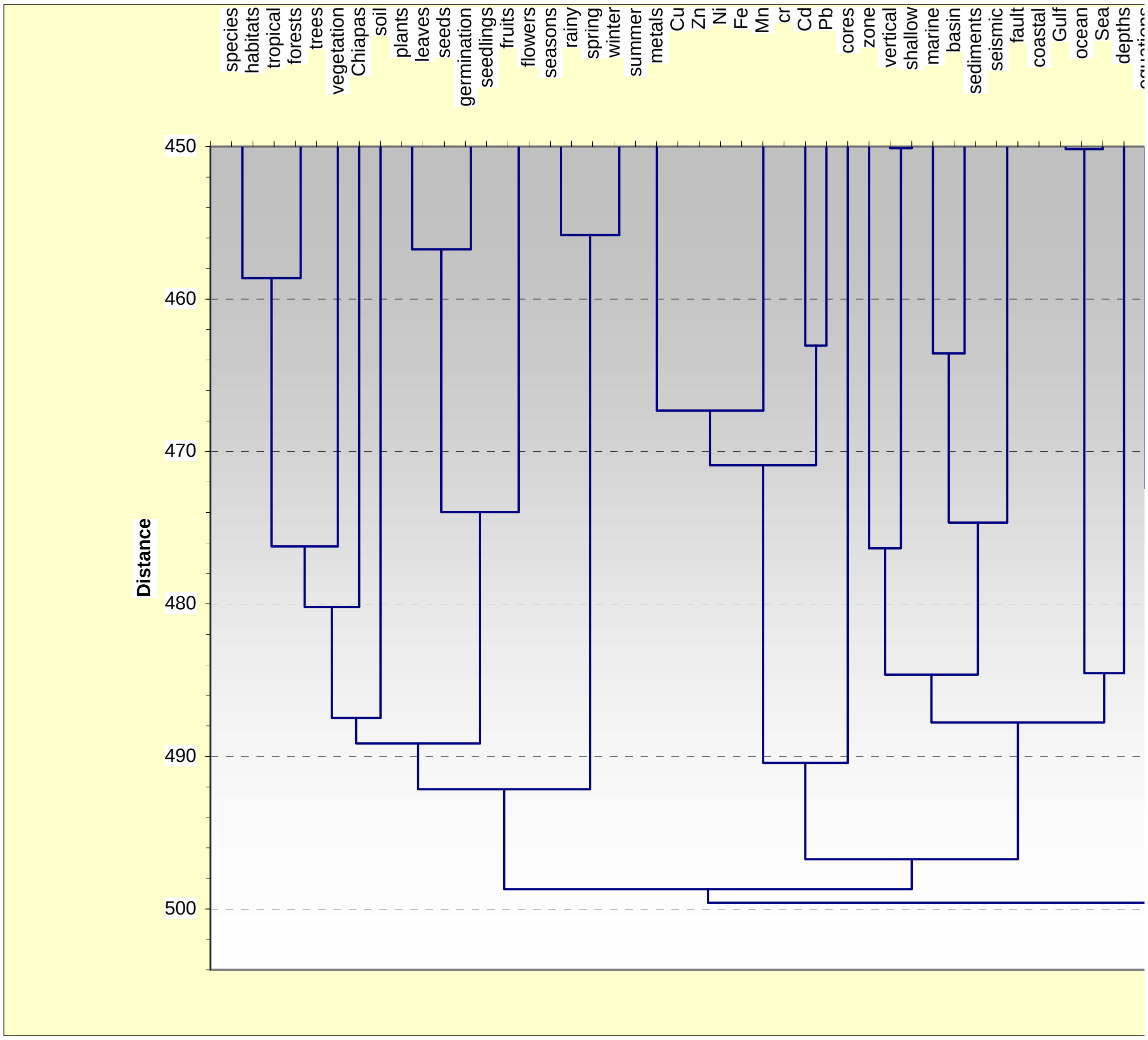
Factor 32 (silicon, Eu, TiO2, luminescence, electronic structure, thin films, fabrication) focuses on thin TiO2:Eu films deposited on silicon substrates, emphasizing electronic structure determination.

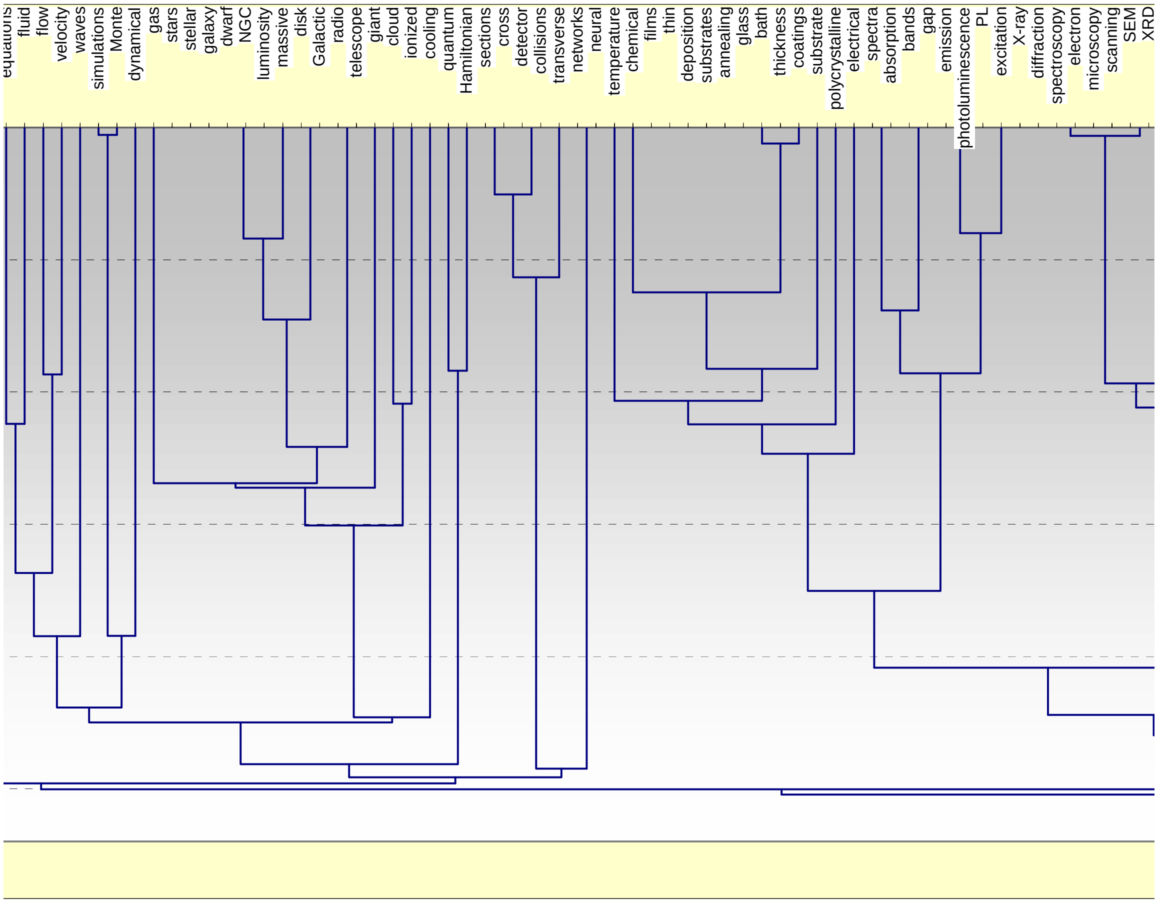
Factor 33 (deforestation, fragmentation, species richness, fragments, rainfall, rainy season, slopes, diversity, soils, canopy, species, highlands, area, conservation, erosion, regeneration) focuses on effects of deforestation and subsequent forest fragmentation on species richness and soil erosion.

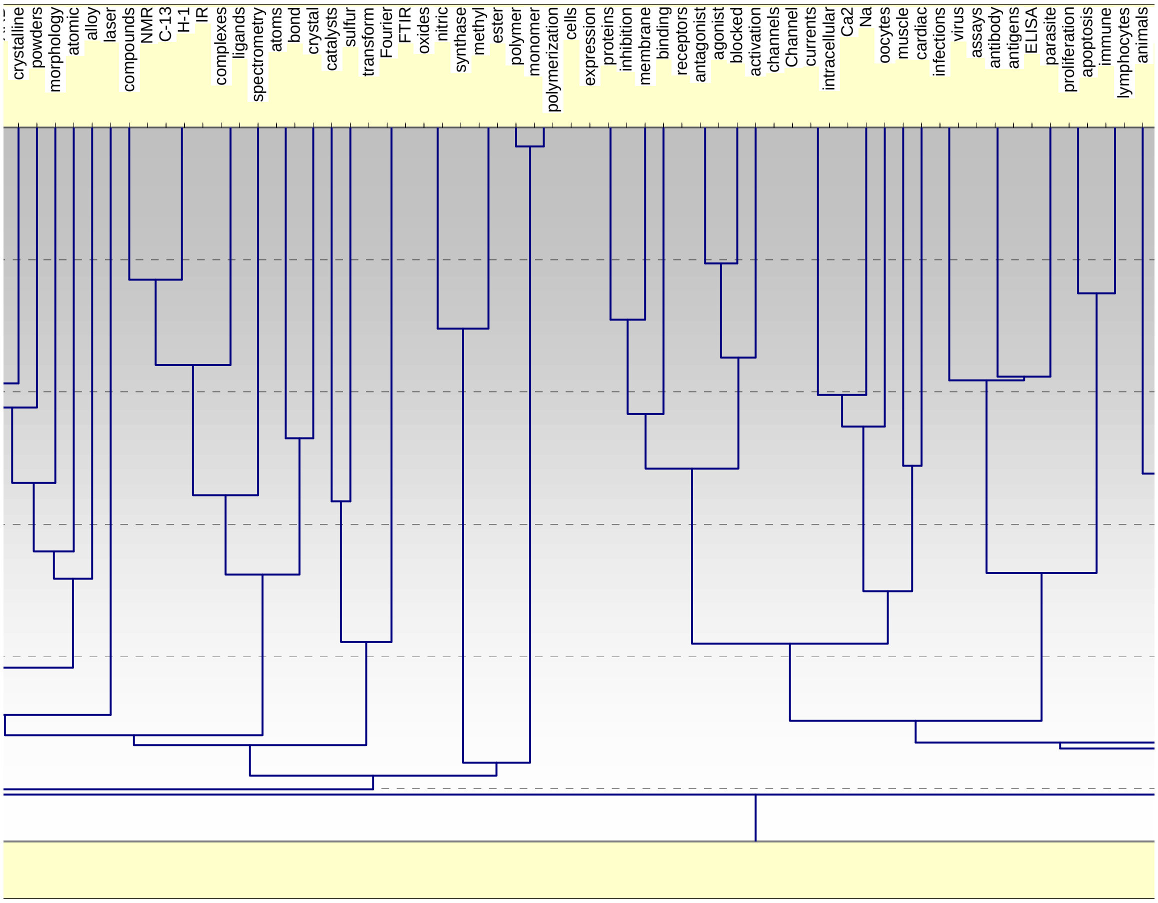
Factor 34 (parasite, host, virus, parasitoids, vitamins, inoculation, viruses, disease, pathogens, competition, larvae) focuses on competition for hosts among parasites, especially for virus0infected species.

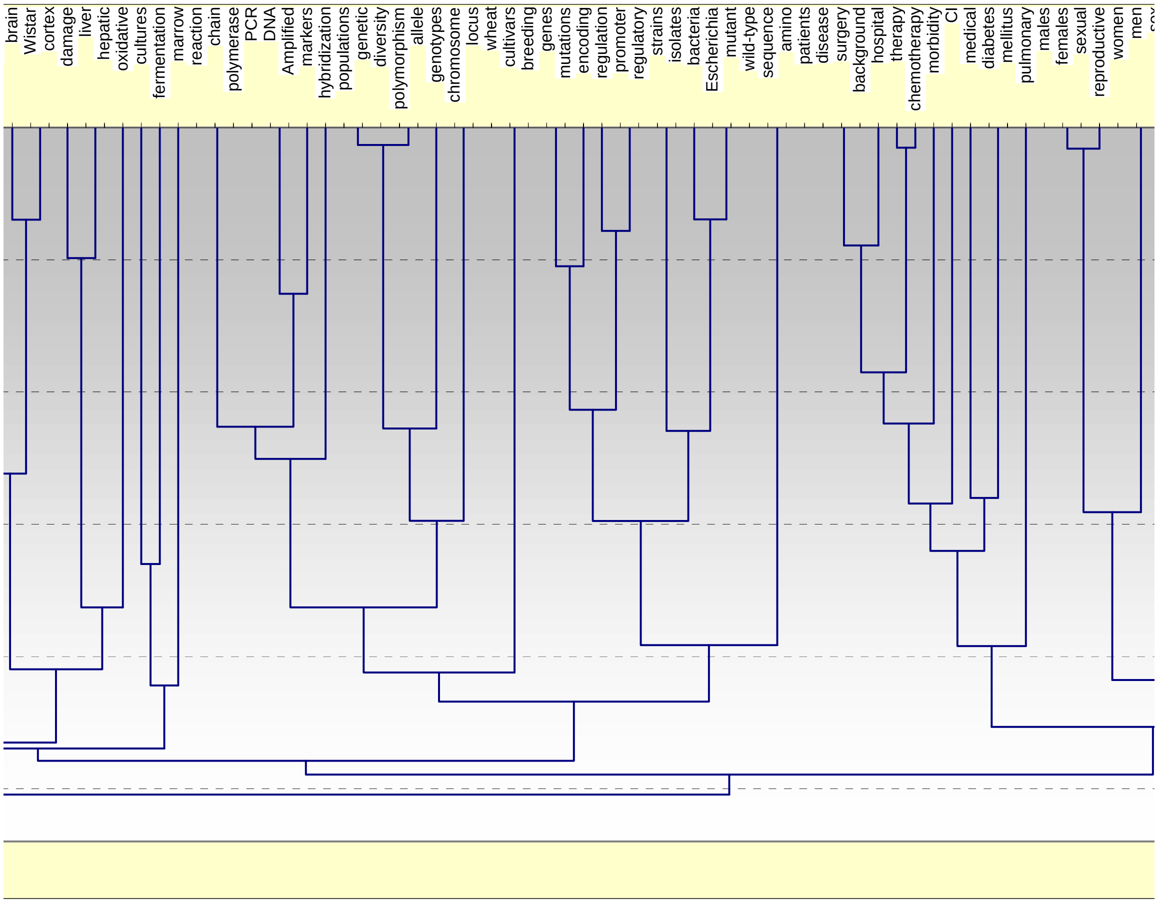
# APPENDIX 4 – WORD DENDROGRAM

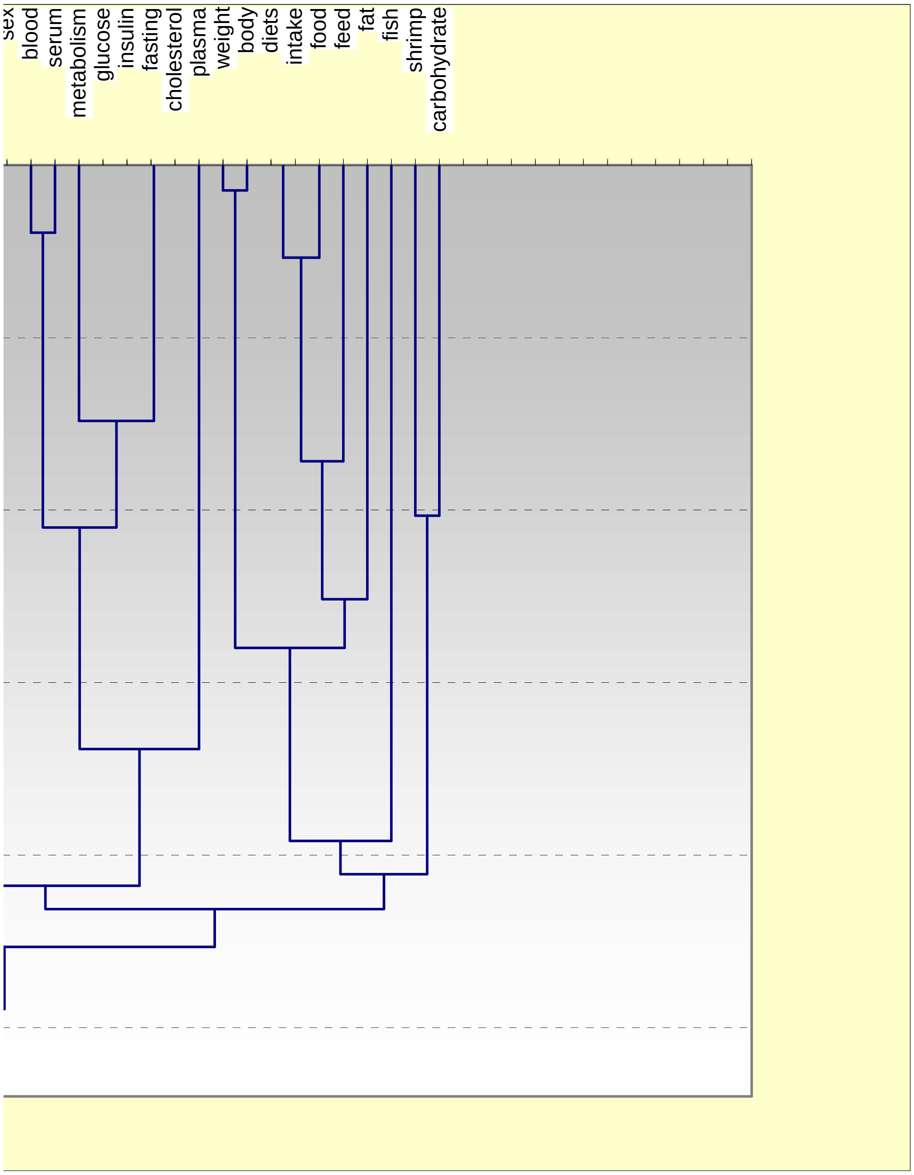
[ADD FROM EXCEL SPREADSHEET]




equations
fluid
flow
velocity
waves
simulations
Monte
dynamical
gas
stars
stellar
galaxy
dwarf
NGC
luminosity
massive
disk
Galactic
radio
telescope
giant
cloud
ionized
cooling
quantum
Hamiltonian
sections
cross
detector
collisions
transverse
networks
neural
temperature
chemical
films
thin
deposition
substrates
annealing
glass
bath
thickness
coatings
substrate
polycrystalline
electrical
spectra
absorption
bands
gap
emission
photoluminescence
PL
excitation
X-ray
diffraction
spectroscopy
electron
microscopy
scanning
SEM
XRD


crystalline
powders
morphology
atomic
alloy
laser
compounds
NMR
C-13
H-1
IR
complexes
ligands
spectrometry
atoms
bond
crystal
catalysts
sulfur
transform
Fourier
FTIR
oxides
nitric
synthase
methyl
ester
polymer
monomer
polymerization
cells
expression
proteins
inhibition
membrane
binding
receptors
antagonist
agonist
blocked
activation
channels
Channel
currents
intracellular
Ca2
Na
oocytes
muscle
cardiac
infections
virus
assays
antibody
antigens
ELISA
parasite
proliferation
apoptosis
immune
lymphocytes
animals


brain
Wistar
cortex
damage
liver
hepatic
oxidative
cultures
fermentation
marrow
reaction
chain
polymerase
PCR
DNA
Amplified
markers
hybridization
populations
genetic
diversity
polymorphism
allele
genotypes
chromosome
locus
wheat
cultivars
breeding
genes
mutations
encoding
regulation
promoter
regulatory
strains
isolates
bacteria
Escherichia
mutant
wild-type
sequence
amino
patients
disease
surgery
background
hospital
therapy
chemotherapy
morbidity
CI
medical
diabetes
mellitus
pulmonary
males
females
sexual
reproductive
women
men
sex


sex
blood
serum
metabolism
glucose
insulin
fasting
cholesterol
plasma
weight
body
diets
intake
food
feed
fat
fish
shrimp
carbohydrate

## APPENDIX 5 – GREEDY STRING TILING METHOD

Greedy String Tiling clustering is a method of grouping text or text character documents (files) by similarity. All documents to be grouped are placed in a database. Each pair of documents is compared by GST, an algorithm originally used to detect plagiarism (42-43), and a similarity score is assigned to the pair. Then, hierarchical aggregation clustering (38-39) is performed on all the documents, using the similarity score for group assignment.

Greedy String Tiling computes the similarity of a pair of documents in two phases. First, all documents to be compared are parsed, and converted into token strings (words or characters). Second, these token strings are compared in pairs for determining the similarity of each pair. During each comparison, the GST algorithm attempts to cover one token string (document) with sub-strings ('tiles') taken from the other string. These sub-strings are not allowed to overlap, resulting in a one to one mapping of tokens. The attribute greedy stems from the fact that the algorithm matches the longest sub-strings first.

A number of similarity metrics can be defined once the tiling is completed. One similarity metric is the percentage of both token strings that is covered. Another similarity metric is the absolute number of shared tokens. A third similarity metric is the mutual information index. Depending on the purpose of the matching, additional weightings can be used for the similarity matrix to increase the ranking precision. For example, if plagiarism is one study objective, additional weighting could be given to shared string length. All similarity metrics have positive and negative features, and the choice of metric is somewhat influenced by the study objectives and the structure of the database.

Once the document similarity matrix has been generated, myriad clustering techniques can be used to produce a classification scheme (taxonomy). In the present study, multi-link hierarchical aggregation was used. Three clustering variants were actually generated, although the extension to other clustering schemes is straight-forward. Single-link, average-link, and complete-link variants are implemented. The variants differ in how the decision of merging to clusters is made. Single-link requires that the similarity of at least two documents is higher than a certain threshold, while complete-link requires that the similarity between all documents in both clusters be higher than a threshold. Average-link requires that the average pair-wise similarity between the documents of both clusters exceed the threshold. For the present study, complete-link appeared to give good results, and was the clustering method used.

## APPENDIX 6 – GREEDY STRING TILING CLUSTERS

(75) Cluster 1) "**c**" (86) "**b**", "**science**" (76) "**v**" (75) "**s**" (20) "**one**" (16) "**system**", "**model**" (15) "**order**" (14) "**d**" (12) "**two**", "**equation**" (11) "**phase**" (10) "**spectral**", "**function**", "**set**", "**space**" (9) "**new**", "**n**", "**based**", "**time**", "**method**", "**pi**", "**energy**", "**nonlinear**", "**design**", "**dimensional**", "**detector**", "**property**" (8) "**science b**", "**b v**" (75) "**c science**" (73) "**d s**", "**second order**" (4) "**gev c**", "**ethyl isopropyl**", "**half wave**", "**spectral ergodicity**", "**confluent retractable**", "**european microbiological**", "**crystallite size**", "**biochemical pathways**", "**property kelley**", "**schrodinger equation**", "**homology z**", "**spot array**", "**coupling strength**", "**model based**", "**effros property**", "**locally connected**", "**n butyl**" (3) "**science b v**" (75) "**c science b**" (73) focuses on general physics, emphasizing mathematical physics, electromagnetism, and gravitational physics.

(26) Cluster 2) "**species**" (47) "**new**" (35) "**mexico**" (19) "**genus**", "**illustrated**" (10) "**n**" (9) "**m**" (8) "**male**", "**dorsal**", "**key**" (7) "**two**", "**sp**", "**genital**", "**state**" (6) "**group**", "**mexican**", "**capsule**" (5) "**collected**", "**region**", "**tribe**", "**habitus**", "**talaromyces**" (4) "**large**", "**leg**", "**d**", "**long**", "**coreidae**", "**reserve**", "**mexicanus**", "**australia**", "**grosshygia**" (3) "**new species**" (28) "**genital capsule**", "**male genital**", "**species genus**" (5) "**two new**", "**new genus**", "**dorsal habitus**" (4) "**n sp**", "**illustrated key**", "**mexico illustrated**" (3) "**male genital capsule**" (5) "**two new species**" (4) "**genus new species**", "**dorsal habitus illustrations**", "**segmented exopod leg**", "**habitus illustrations drawings**", "**drawings male genital**", "**habitus antennal segments**", "**new species species**", "**dorsal habitus antennal**", "**mexico new species**", "**illustrations drawings male**", "**species species genus**", "**genital capsule parameres**", "**new genus new**", "**alloophorus robustus lake**" (2) focuses on species of insects and animals.

(25) Cluster 3) "**patients**" (150) "**treatment**" (39) "**years**", "**disease**" (35) "**chemotherapy**" (31) "**median**" (24) "**toxicity**" (23) "**age**", "**survival**", "**months**" (22) "**one**", "**response**" (21) "**follow**" (20) "**surgery**", "**cases**", "**treated**" (18) "**group**" (17) "**complications**", "**mean**", "**cancer**" (16) "**surgical**", "**outcome**", "**95**" (15) "**two**", "**methods**", "**four**", "**overall**", "**carcinoma**", "**secondary**" (14) "**three**", "**20**", "**complete**", "**therapy**", "**doses**", "**received**", "**breast**" (13) "**free**", "**100**", "**advanced**", "**ifn**" (12) "**patients treated**" (11) "**overall survival**", "**95 ci**", "**patients received**" (8) "**mg m**", "**breast carcinoma**", "**free survival**", "**follicular lymphoma**", "**long term**" (7) "**cervical carcinoma**", "**antibiotic treatment**", "**secondary peritonitis**", "**locally advanced**", "**conduction disturbances**" (6) "**median follow**", "**mean age**", "**cd 20**", "**toxicity mild**", "**event free**", "**years age**", "**two patients**", "**four patients**", "**rectal cancer**", "**chemotherapy ifn**", "**refractory follicular**" (5) "**event free survival**", "**refractory follicular lymphoma**" (5) "**patients refractory follicular**", "**connective tissue disease**", "**95 confidence interval**", "**lymphoma heavily treated**", "**locally advanced cervical**", "**anti cd 20**", "**follicular lymphoma heavily**", "**mixed connective tissue**" (4) focuses on clinical studies of chronic diseases, mainly cancer.

(19) Cluster 4) "**films**" (81) "**temperature**" (23) "**thin**" (18) "**substrates**", "**deposited**" (17) "**cdte**" (16) "**concentration**", "**solution**", "**resistivity**" (15) "**c**", "**glass**", "**x**" (14)

"**degreesc**" (13) "**structure**", "**cm**", "**nm**", "**cd**" (12) "**band**", "**tio2**" (11) "**surface**", "**temperatures**", "**photoluminescence**", "**film**", "**substrate**", "**grown**" (10) "**spectra**", "**range**", "**size**" (9) "**low**", "**emission**", "**science**", "**optical**", "**gap**", "**ev**", "**electrical**", "**room**", "**pl**" (8) "**thin films**" (18) "**films deposited**", "**glass substrates**" (12) "**band gap**", "**c science**" (8) "**omega cm**", "**room temperature**" (7) "**x ray**", "**films grown**" (6) "**pl signal**", "**surface morphology**", "**sol gel**", "**c american**", "**deposited silicon**", "**starting solution**", "**grain size**" (5) "**force microscopy**", "**films sol**", "**coming glass**", "**science b**", "**b v**", "**substrate temperature**", "**atomic force**", "**d f**" (4) "**science b v**", "**c science b**", "**thin films deposited**", "**atomic force microscopy**", "**films sol gel**" (4) "**concentration starting solution**", "**zno thin films**", "**x omega cm**", "**characteristics films deposited**", "**films deposited silicon**", "**deposited silicon wafers**", "**d f transition**", "**signal d f**", "**x ray diffraction**", "**c american physics**", "**pl signal d**", "**coming glass substrates**", "**thin films sol**", "**transition electronic structure**" (3) focuses on thin films, especially deposition and properties.

(17) Cluster 5) "**c**" (22) "**experimental**" (21) "**v**" (19) "**b**" (18) "**science**" (17) "**spectra**", "**model**" (15) "**phase**" (11) "**two**" (10) "**laser**", "**calculations**" (9) "**ii**", "**data**", "**method**", "**films**", "**pm3**" (8) "**one**", "**surface**", "**s**", "**raman**", "**complex**", "**hf**", "**oxide**" (7) "**n**", "**growth**", "**order**", "**carried**", "**solid**", "**gas**", "**film**", "**systems**", "**r**", "**mode**", "**infrared**", "**fourier**", "**transform**", "**pore**" (6) "**c science**", "**science b**", "**b v**" (17) "**fourier transform**" (6) "**raman spectra**", "**experimental data**" (5) "**spectra solid**", "**contact mode**", "**gas phase**" (4) "**transform infrared**", "**solid phase**", "**cd ii**", "**hf 31g**", "**pb ii**", "**ammonium jarosite**", "**terephthalic acid**", "**pm3 tm**", "**calculations carried**" (3) "**science b v**", "**c science b**" (17) "**fourier transform infrared**" (3) focuses on measurement of material and compound spectra, mainly IR and Raman.

(17) Cluster 6) "**films**" (30) "**temperature**" (21) "**c**" (19) "**b**" (18) "**v**", "**science**" (16) "**k**", "**thin**" (12) "**low**" (10) "**omega**", "**brane**" (9) "**deposition**" (8) "**composition**", "**potential**", "**high**", "**chemical**", "**temperatures**", "**quantum**", "**cdte**" (7) "**growth**", "**oxygen**", "**single**", "**film**", "**substrate**", "**room**" (6) "**surface**", "**conditions**", "**process**", "**cell**", "**laser**", "**si**", "**co**", "**cr**", "**polycrystalline**", "**grain**", "**j**", "**deposited**", "**cigs**", "**defect**" (5) "**science b**", "**b v**", "**c science**" (16) "**thin films**" (12) "**omega k**" (7) "**room temperature**" (6) "**chemical composition**" (4) "**barrier height**", "**polycrystalline si**", "**cigs films**", "**cn gate**", "**cdte thin**", "**delta omega**", "**low temperature**", "**k omega**", "**pulse quantum**", "**magnetic field**" (3) "**c science b**", "**science b v**" (16) "**omega k omega**", "**k omega k**", "**cdte thin films**" (3) focuses on deposition and growth of thin films.

(16) Cluster 7) "**cross**" (38) "**p**", "**section**" (23) "**data**", "**gev**" (22) "**jet**" (19) "**momentum**", "**measurement**" (17) "**transverse**" (16) "**q**", "**sections**" (15) "**proton**", "**qcd**" (13) "**alpha**", "**t**" (12) "**measured**", "**d**", "**w**", "**ratio**", "**roots**" (11) "**z**", "**c**", "**mass**", "**detector**", "**inclusive**", "**diffractive**" (10) "**range**", "**function**", "**production**", "**leading**" (9) "**order**", "**energy**", "**photon**", "**agreement**", "**fermilab**", "**collisions**" (8) "**cross section**" (23) "**cross sections**" (15) "**transverse momentum**" (13) "**fermilab tevatron**" (7) "**c science**", "**differential cross**", "**science b**", "**1800 gev**", "**collisions roots**", "**b v**", "**next leading**", "**leading order**" (6) "**deep inelastic**", "**roots 630**", "**three**

**jet**", "**jet cross**" (5) "**next leading order**", "**c science b**", "**science b v**" (6) "**leading order qcd**", "**fermilab tevatron collider**", "**p bar collisions**", "**differential cross sections**", "**p p bar**" (4) focuses on cross section measurements of high energy beams, mainly proton.

(16) Cluster 8) "**grating**" (8) "**one**", "**two**", "**field**", "**dimensional**" (7) "**fields**" (6) "**i**", "**scattering**", "**opt**" (5) "**range**", "**method**", "**medium**", "**structure**", "**coherent**", "**space**", "**light**" (4) "**source**", "**new**", "**18**", "**surface**", "**order**", "**techniques**", "**numerical**", "**functions**", "**case**", "**surfaces**", "**waveguide**", "**fixed**", "**point**", "**j**", "**conserved**", "**rays**", "**analogs**", "**fractal**", "**comment**", "**definitions**", "**080**" (3) "**optical america**", "**c optical**" (16) "**two dimensional**", "**j opt**", "**conserved rays**", "**field c**", "**fixed grating**" (3) "**soc 18**", "**three dimensional**", "**opt lett**", "**scattering one**", "**coherent monochromatic**", "**fractional talbot**", "**finite grating**", "**third order**", "**america ocis**", "**rigorous numerical**", "**ocis codes**", "**material response**", "**one dimensional**", "**light scattering**", "**space charge**", "**free space**", "**18 902**", "**opt soc**", "**electric field**", "**i comment**", "**point source**" (2) "**c optical america**" (16) "**field c optical**" (3) "**america ocis codes**", "**optical america ocis**", "**j opt soc**", "**electric field c**", "**opt soc 18**" (2) focuses on light scattering from surfaces, emphasizing gratings.

(15) Cluster 9) "**patients**" (85) "**reading**" (15) "**one**", "**months**" (14) "**two**", "**diastolic**", "**stroke**" (12) "**12**", "**group**", "**methods**" (11) "**portal**", "**hypertension**", "**surgery**", "**ischemic**" (10) "**patient**", "**objective**", "**delivery**", "**ventricular**", "**sphincter**" (9) "**age**", "**time**", "**risk**", "**external**", "**seizures**", "**symptoms**", "**left**", "**dysfunction**", "**ultrasound**", "**neck**", "**progressors**" (8) "**vaginal delivery**", "**external ultrasound**" (6) "**ventricular diastolic**", "**risk factors**", "**distal pancreatectomy**", "**ischemic stroke**", "**venous sampling**", "**two patients**", "**left ventricular**", "**scuba sky**" (5) "**ventricular dysfunction**", "**diastolic function**", "**selective venous**", "**neck stiffness**", "**stapes surgery**", "**diabetic patients**", "**systemic arterial**", "**rebleeding rate**", "**operative mortality**", "**atrial fibrillation**", "**portal hypertension**", "**tissue retraction**", "**degree tissue**", "**anal sphincter**", "**arterial hypertension**" (4) "**systemic arterial hypertension**", "**selective venous sampling**", "**degree tissue retraction**" (4) "**diastolic blood pressure**", "**left ventricular diastolic**", "**ventricular diastolic function**", "**flow preserving procedures**", "**blood flow preserving**", "**portal blood flow**" (3) focuses on clinical medical research, emphasizing cardiovascular problems.

(15) Cluster 10) "**patients**" (54) "**p**" (49) "**group**" (43) "**n**", "**groups**" (26) "**vs**" (23) "**mg**" (22) "**treatment**", "**months**" (18) "**days**", "**years**", "**placebo**" (15) "**001**", "**baseline**" (14) "**mk**", "**insulin**", "**aln**", "**869**" (13) "**daily**" (12) "**increased**" (11) "**period**", "**12**", "**05**", "**acute**", "**month**" (10) "**similar**", "**decreased**", "**dose**", "**24**", "**28**", "**methods**", "**equal**", "**bone**", "**cisplatin**", "**emesis**" (9) "**p 001**", "**mk 869**" (13) "**p 05**", "**mg po**", "**double blind**" (7) "**vs p**", "**latin america**", "**p equal**", "**group n**" (6) "**pulmonary tuberculosis**", "**stent group**", "**inhaled beclomethasone**", "**insulin sensitivity**" (5) "**n pufa**", "**days group**", "**mellitus patients**", "**ii iii**", "**month follow**", "**group i**", "**5ht antagonist**", "**delayed emesis**", "**pre cisplatin**", "**diabetes mellitus**", "**antagonist dexamethasone**", "**gi aes**", "**type diabetes**", "**24 months**", "**intra oesophageal**" (4) "**type diabetes**

**mellitus**", "**diabetes mellitus patients**" (4) focuses on clinical medical research, emphasizing diabetes and related illnesses.

(15) Cluster 11) "**patients**" (129) "**disease**" (48) "**p**" (35) "**acyclovir**" (25) "**hsv**" (22) "**95**" (20) "**group**", "**years**", "**ci**" (19) "**treatment**" (16) "**days**", "**controls**", "**hd**" (14) "**vs**", "**methods**", "**cases**", "**therapy**", "**ratio**" (13) "**12**", "**months**", "**objective**" (12) "**age**", "**s**", "**mortality**", "**infection**", "**treated**", "**kg**", "**symptoms**", "**mean**", "**iga**", "**joas**", "**aoas**", "**saliva**", "**sibutramine**" (11) "**hsv disease**" (13) "**95 ci**", "**hd acyclovir**" (10) "**odds ratio**" (9) "**controls p**", "**om 85**", "**85 bv**" (7) "**ci 95**" (6) "**41 patients**", "**mg kg**", "**neonatal hsv**", "**p 001**", "**kg d**", "**antiviral therapy**", "**patients treated**", "**open shunt**" (5) "**om 85 bv**" (7) "**mg kg d**" (5) "**patients open shunt**", "**iga protein ratio**", "**neonatal hsv disease**" (4) "**chronic liver disease**", "**susceptible children 12**", "**years odds ratio**", "**confidence interval ci**", "**treated hd acyclovir**", "**p s group**", "**initiation antiviral therapy**", "**85 bv group**", "**disseminated hsv disease**", "**95 confidence interval**", "**mg dl ci**", "**patients cns disease**", "**disease diabetes mellitus**", "**dl ci 95**" (3) focuses on infectious diseases.

(13) Cluster 12) "**scale**", "**inflation**" (8) "**growth**", "**field**", "**systems**" (7) "**extra**", "**dimensions**" (6) "**new**", "**years**", "**matrix**",  "**branes**" (5) "**large**", "**stable**", "**control**", "**high**",  "**concentration**", "**pressure**", "**stability**", "**universe**",, "**theories**",  "**extra dimensions**" (4) "**affleck dine**", (3) "**steiner triple systems**", "**extra dimensions stabilized**", "**size extra dimensions**",  "**stable d3 branes**", "**affleck dine field**" (2) focuses on cosmology, emphasizing inflation.

(13) Cluster 13) "**patients**" (76) "**p**" (72) "**001**", "**hla**", "**drb1**" (21) "**controls**", "**sle**" (20) "**pi**" (18) "**mortality**" (17) "**disease**" (15) "**alleles**" (14) "**vs**", "**05**" (13) "**mexican**" (12) "**decreased**", "**treatment**", "**d**", "**38**", "**year**", "**factors**", "**67**" (11) "**age**", "**c**", "**group**", "**cases**", "**allele**" (10) "**variables**", "**increased**", "**pulmonary**", "**pc**", "**tnf**", "**copd**" (9) "**control**", "**patient**", "**risk**", "**class**", "**ln**", "**dqb1**", "**exercise**", "**dyspnea**" (8) "**p 001**" (21) "**p 05**" (10) "**38 patients**", "**pi treatment**" (8) "**ln patients**", "**class ii**", "**year pi**", "**ki 67**" (6) "**drb1 0301**", "**p 01**", "**p c**", "**risk factors**" (5) "**odds ratio**", "**drb1 0802**", "**hla drb1**", "**mestizo patients**", "**sle patients**", "**genetic factors**", "**groups p**", "**p 02**", "**95 ci**", "**pf patients**", "**mexican mestizo**", "**mexican patients**", "**tnf alpha**" (4) "**polymerase chain reaction**", "**lupus erythematosus sle**", "**tnf2 308 d**", "**year pi treatment**", "**obstructive pulmonary disease**", "**chronic obstructive pulmonary**", "**mexican mestizo patients**", "**sigma vs controls**", "**mortality risk factors**", "**index p 001**", "**class ii alleles**", "**takayasu s arteritis**", "**systemic lupus erythematosus**", "**multiple logistic regression**" (3) focuses on clinical medical research, emphasizing autoimmune and inflammatory diseases.

(13) Cluster 14) "**cell**" (78) "**cells**" (50) "**apoptosis**" (26) "**p53**" (25) "**host**", "**death**", "**csf**" (16) "**apoptotic**" (15) "**proliferation**", "**dna**", "**gm**" (14) "**lines**" (13) "**protein**" (11) "**survival**", "**human**", "**bcl**", "**mdm4**" (10) "**surface**", "**factor**", "**sf9**", "**topoisomerase**" (9) "**dependent**", "**activity**", "**cycle**", "**cultures**", "**cd34**", "**scl**" (8) "**normal**", "**ii**", "**cellular**", "**interaction**", "**response**", "**plant**" (7) "**gm csf**" (14) "**cell death**" (13) "**cell lines**" (12) "**topoisomerase ii**", "**cell proliferation**" (7) "**justicia spicigera**", "**cell line**",

"**cell cycle**" (6) "**host cell**", "**sf9 cells**" (5) "**32d cells**", "**cd34 cells**", "**cell surface**", "**flow cytometry**" (4) "**membrane blebbing**", "**hl 60**", "**induces apoptosis**", "**transcription factor**", "**mdm2 mdm4**", "**t 514**", "**response gm**", "**cells mouse**", "**plant extract**", "**extract justicia**", "**dna degradation**", "**sf gm**", "**host cells**", "**sialic acid**", "**cell viability**", "**sub g1**" (3) "**sf gm csf**", "**response gm csf**", "**extract justicia spicigera**" (3) focuses on cell biology related to human illness.

(13) Cluster 15) "**species**" (109) "**new**" (39) "**p**" (18) "**mexico**" (15) "**region**" (13) "**habitat**" (11) "**based**", "**body**", "**forests**" (9) "**parts**", "**area**", "**e**", "**lists**" (8) "**genus**", "**taxa**", "**families**" (7) "**tropical**", "**richness**", "**southern**", "**cactus**", "**conservation**", "**pine**" (6) "**three**", "**cells**", "**system**", "**collected**", "**first**", "**plant**", "**diversity**", "**endemic**", "**climate**", "**ecological**", "**chiapas**", "**san**", "**specimens**", "**desert**", "**fauna**", "**deciduous**", "**huizache**" (5) "**new species**" (28) "**species p**", "**species lists**" (7) "**new genus**", "**species pine**" (5) "**chihuahuan desert**", "**tropical deciduous**", "**cactus species**", "**erect system**", "**species richness**", "**eastern pacific**", "**body parts**" (4) "**mum long**", "**epistylid species**", "**genus new**", "**apparent habitat**", "**body region**", "**monosporangia equal**", "**oak forests**", "**huizache area**", "**restricted species**", "**pine oak**", "**omnipresent taxa**" (3) "**new species p**" (7) "**new species pine**" (5) "**new genus new**", "**pine oak forests**", "**species pine oak**" (3) focuses on new fauna species.

(12) Cluster 16) "**controller**" (31) "**control**" (21) "**observer**" (14) "**sliding**" (13) "**tracking**" (12) "**state**", "**mode**", "**feedback**" (11) "**system**", "**time**", "**parameters**", "**output**" (10) "**based**", "**input**" (9) "**c**", "**model**", "**design**", "**loop**", "**reference**" (8) "**order**", "**paper**" (7) "**simulations**", "**reduced**", "**systems**", "**stability**", "**convergence**", "**closed**" (6) "**sliding mode**" (10) "**closed loop**" (6) "**numerical simulations**", "**c science**" (5) "**reduced order**", "**finite time**" (4) "**perfect tracking**", "**feedback control**", "**reference trajectories**", "**controller observer**", "**sliding surface**", "**varying parameters**", "**input signals**", "**passivity based**", "**mode control**", "**sufficient conditions**", "**plant parameters**", "**dynamic sliding**", "**pi control**", "**c sons**", "**asymptotic stability**", "**spatially varying**" (3) "**sliding mode control**" (3) focuses on feedback control of complex systems.

(12) Cluster 17) "**species**" (54) "**sp**" (22) "**m**" (15) "**new**" (14) "**helminth**" (13) "**mexico**", "**freshwater**", "**fishes**" (9) "**monstrilla**", "**specimens**", "**helminths**" (7) "**collected**", "**caribbean**", "**recorded**", "**first**", "**taxonomic**", "**fauna**" (6) "**one**", "**gen**", "**fish**", "**records**", "**monstrilloid**", "**cymbasoma**", "**original**", "**parasites**", "**nicaragua**", "**saccocoelioides**", "**ascocotyle**" (5) "**new species**", "**helminth species**", "**gen sp**" (5) "**monstrilloid copepods**", "**saccocoelioides sp**", "**fish species**", "**freshwater fishes**" (4) "**m bondae**", "**helminth parasites**", "**balsas river**", "**freshwater fish**", "**ascocotyle phagicola**", "**digenean species**", "**22 species**", "**species monstrilla**", "**helminth fauna**", "**sp nov**" (3) "**freshwater fish species**" (3) "**south atlantic autonomous**", "**fishes southeastern mexico**", "**atlantic autonomous region**", "**new species monstrilla**", "**freshwater fishes southeastern**", "**balsas river drainage**", "**tapeworm bothriocephalus acheilognathi**", "**autonomous region nicaragua**", "**garter snake thamnophis**", "**fishes balsas river**", "**species freshwater fishes**", "**species similar m**",

"**new species similar**", "**fish species south**", "**species south atlantic**" (2) focuses on new species of fish.

(12) Cluster 18) "**energy**" (21) "**phase**" (16) "**b**", "**v**", "**c**", "**science**" (12) "**space**" (10) "**two**", "**model**", "**stable**" (7) "**one**", "**potential**", "**small**", "**method**", "**amplitude**", "**systems**" (6) "**number**", "**time**", "**structure**", "**critical**", "**wave**", "**gravity**", "**global**", "**discrete**", "**string**" (5) "**b v**", "**c science**", "**science b**" (12) "**leu enkephalin**" (4) "**de sitter**", "**discrete fresnel**", "**phase space**" (3) "**c science b**", "**science b v**" (12) "**page phase transitions**", "**systems c science**", "**discrete fresnel integral**", "**comparison wave front**", "**total monopole energy**", "**global energy minimum**", "**conformations leu enkephalin**", "**hawking page phase**", "**low energy conformations**" (2) focuses on energy methods for optimization.

(12) Cluster 19) "**soil**" (105) "**c**" (38) "**erosion**" (21) "**biomass**" (20) "**n**", "**model**" (19) "**years**" (18) "**increased**", "**straw**" (15) "**cm**", "**fine**" (14) "**14**", "**root**", "**kg**" (13) "**co2**" (12) "**soils**", "**drained**" (11) "**science**", "**dynamics**", "**microbial**" (10) "**production**", "**loss**", "**organic**", "**tdf**" (9) "**b**", "**three**", "**v**", "**total**", "**p**", "**mexico**", "**productivity**", "**glucose**" (8) "**c 14**" (13) "**soil erosion**" (12) "**fine root**" (10) "**c science**", "**microbial biomass**" (9) "**b v**", "**science b**" (8) "**soil loss**", "**drained years**" (7) "**tdf pasture**", "**pb straw**", "**soil drained**" (6) "**kg soil**", "**root biomass**", "**cm soil**", "**co2 production**", "**co2 c**", "**mulch layer**" (5) "**science b v**" (8) "**c science b**" (7) "**fine root biomass**" (5) "**co2 c 14**" (4) "**live fine root**", "**total soil loss**", "**c 14 labelled**", "**first cm soil**", "**former lake texcoco**", "**plant residue mulch**", "**residue mulch layer**", "**c 14 c**", "**14 labelled glucose**", "**soil drained years**", "**soil former lake**", "**ctb straw incorporated**", "**straw burned treatment**", "**microbial biomass c**" (3) focuses on soil erosion and dynamics.

(12) Cluster 20) "**95**" (48) "**ci**" (38) "**women**" (35) "**treatment**" (24) "**risk**" (20) "**years**" (19) "**model**" (18) "**mexico**" (15) "**new**", "**factors**" (14) "**patients**", "**abuse**" (13) "**age**" (12) "**vs**", "**higher**", "**symptoms**", "**sexual**", "**cancer**", "**hsv**" (11) "**population**", "**rr**", "**cervical**", "**nurses**" (10) "**pulmonary**", "**cl**", "**background**", "**care**", "**tuberculosis**" (9) "**sample**", "**increased**", "**n**", "**mortality**", "**survival**", "**infection**", "**methods**", "**ratio**", "**23**", "**city**", "**hpv**", "**eclampsia**" (8) "**95 ci**" (32) "**new model**" (12) "**cervical cancer**" (10) "**mexico city**", "**rr 95**", "**risk factors**" (8) "**physical sexual**", "**95 cl**" (7) "**population based**", "**treatment seeking**" (6) "**confidence interval**", "**drug resistance**", "**hpv vaccine**", "**40 years**", "**sexual abuse**", "**years old**", "**standard model**", "**95 confidence**" (5) "**nurses aides**", "**higher risk**", "**eclampsia eclampsia**", "**hazards ratio**", "**abuse adulthood**", "**emotional abuse**", "**pre eclampsia**", "**cessation breastfeeding**", "**hsv seroprevalence**", "**early cessation**" (4) "**rr 95 ci**" (7) "**95 confidence interval**", "**physical sexual abuse**", "**40 years old**" (5) "**pre eclampsia eclampsia**" (4) "**odds ratio 95**", "**confidence interval ci**", "**risk factors cervical**", "**early cessation breastfeeding**", "**population based hsv**", "**factors cervical cancer**", "**women mexico city**" (3) focuses on clinical research into women’s health problems, emphasizing cervical cancer.

(12) Cluster 21) "**degreesc**" (34) "**thin**" (20) "**films**" (19) "**nm**", "**omega**" (14) "**layer**" (13) "**phase**" (12) "**band**", "**square**", "**substrates**", "**cds**" (11) "**x**", "**film**", "**100**", "**deposited**", "**foils**" (10) "**ray**", "**thickness**", "**cd**" (9) "**25**", "**min**", "**coated**",

"**resistance**", "**sheet**", "**annealed**", "**coatings**", "**cdo**" (8) "**range**", "**500**", "**crystalline**", "**ev**" (7) "**thin films**" (11) "**x ray**" (9) "**sheet resistance**" (8) "**500 degreesc**" (7) "**ray diffraction**", "**omega cm**" (6) "**thin film**", "**chemical bath**", "**omega square**", "**300 degreesc**", "**cds thin**", "**100 nm**" (5) "**electrical conductivity**", "**cdo layer**", "**370 degreesc**", "**yellow band**" (4) "**x ray diffraction**" (6) "**cds thin film**" (4) "**copper sulfide thin**", "**thin layer cdo**", "**sulfide thin films**", "**c science b**", "**100 nm thickness**", "**thin films 75**", "**science b v**" (3) focuses on deposition, growth, and properties of thin films.

(11) Cluster 22) "**patients**" (106) "**dialysis**" (25) "**disease**" (19) "**nail**" (17) "**fabry**", "**renal**", "**anemia**" (14) "**end**" (13) "**lower**", "**cri**" (11) "**levels**", "**rv**", "**hematocrit**", "**onychomycosis**" (10) "**alterations**", "**body**", "**mean**", "**viral**", "**load**", "**ma**", "**pgr**" (9) "**increased**", "**chronic**", "**year**", "**serum**", "**ventricular**", "**diastolic**", "**venous**", "**rhuepo**", "**er**" (8) "**type**", "**higher**", "**peritoneal**", "**hct**" (7) "**renal disease**" (9) "**fabry patients**" (8) "**peritoneal dialysis**" (7) "**end stage**", "**nail alterations**", "**left ventricular**", "**stage renal**", "**fabry disease**", "**viral load**" (6) "**united states**", "**nyha class**", "**rv ischemia**", "**serum creatinine**" (5) "**end diastolic**", "**load patients**", "**venous insufficiency**", "**management anemia**", "**body segment**", "**patients chronic**", "**mass index**", "**history hemodialysis**", "**patients initiated**", "**nine patients**", "**initiated dialysis**", "**patients obesity**", "**hematocrit levels**" (4) "**stage renal disease**", "**end stage renal**" (6) "**fabry patients initiated**", "**patients initiated dialysis**" (4) "**glomerular filtration rate**", "**body mass index**", "**peritoneal dialysis history**", "**dialysis history hemodialysis**", "**patients end stage**", "**non diabetic controls**", "**viral load patients**", "**type dm patients**" (3) focuses on clinical research in renal disease, emphasizing relationship with cardiovascular disease.

(11) Cluster 23) "**p**" (18) "**c**", "**molecules**" (14) "**density**", "**state**" (13) "**molecular**" (12) "**b**", "**local**", "**calculations**" (11) "**science**", "**theory**", "**charge**" (9) "**interaction**", "**energy**", "**functional**", "**sensitivity**", "**electronic**", "**global**" (8) "**v**", "**complexes**", "**h**", "**reaction**", "**functions**", "**atomic**", "**charges**", "**fukui**", "**reactivity**", "**molecule**" (7) "**chemical**", "**structure**", "**states**", "**x**", "**transfer**", "**dipole**", "**b3lyp**", "**dioxide**", "**net**", "**hartree**", "**fock**" (6) "**c science**" (9) "**density functional**", "**global local**" (8) "**fukui functions**", "**b v**", "**science b**", "**functional theory**" (7) "**hartree fock**" (6) "**atomic charges**", "**charge transfer**" (5) "**sensitivity coefficients**", "**charges global**", "**p p**", "**ct complexes**", "**coefficients molecular**", "**electronic structure**", "**net atomic**", "**molecular energy**", "**x ray**", "**obtaining sensitivity**", "**charge sensitivity**", "**local softness**", "**local hardness**", "**hardness global**", "**energy net**", "**softness fukui**" (4) "**science b v**", "**c science b**", "**density functional theory**" (7) "**atomic charges global**", "**molecular energy net**", "**hardness global local**", "**net atomic charges**", "**global local softness**", "**global local hardness**", "**softness fukui functions**", "**obtaining sensitivity coefficients**", "**charges global local**", "**local hardness global**", "**coefficients molecular energy**", "**sensitivity coefficients molecular**", "**energy net atomic**", "**local softness fukui**" (4) focuses on electronic structure and reactivity ab initio calculations.

(11) Cluster 24) "**emission**" (25) "**optical**" (17) "**absorption**" (16) "**nm**" (13) "**ions**", "**fluorescence**" (11) "**doped**", "**matrix**", "**radiative**" (10) "**b**", "**v**", "**c**", "**science**", "**samples**" (9) "**red**", "**properties**", "**crystal**", "**excitation**" (7) "**g**", "**cross**", "**h**", "**ion**",

"**laser**", "**pmma**" (6) "**sample**", "**f**", "**spectra**", "**content**", "**energy**", "**transition**", "**nd3**", "**host**" (5) "**c science**" (9) "**science b**", "**b v**" (8) "**optical absorption**" (7) "**emission cross**" (5) "**cross section**", "**optical gain**", "**judd ofelt**", "**h g**" (4) "**absorption emission**", "**x ray**", "**low temperature**", "**glassy matrix**", "**nd3 ions**", "**quenched samples**", "**g h**", "**energy transfer**" (3) "**science b v**", "**c science b**" (8) "**emission cross section**" (3) "**cross section sigma**", "**judd ofelt theory**", "**plastic optical fibers**", "**h g h**", "**crystal c science**", "**emission cross sections**", "**radiative decay time**", "**sections optical gain**", "**viz radiative transition**", "**ions c science**", "**lambda 600 nm**", "**fluoroborophosphate glassy matrix**", "**g h g**", "**stimulated emission cross**", "**cross sections optical**", "**radiative energy transfer**" (2) focuses on the emission and excitation spectra of elements and compounds, emphasizing optical spectra.

(11) Cluster 25) "**method**" (30) "**based**" (13) "**h**", "**d**" (7) "**new**", "**c**" (6) "**system**", "**images**" (5) "**n**", "**model**", "**numerical**", "**algorithm**", "**feces**" (4) "**v**", "**spectral**", "**s**", "**function**", "**methods**", "**paper**", "**terms**", "**gamma**", "**turbulence**", "**recovery**", "**classical**", "**gravity**", "**accuracy**", "**shape**", "**vertical**", "**element**", "**compute**", "**scidar**", "**quadrangles**", "**relocation**", "**stereo**" (3) "**method based**" (7) "**n h**", "**c n**" (4) "**d shape**", "**spectral element**", "**v h**", "**new method**" (3) "**collecting feces**", "**numerical method**", "**classical spectral**", "**vertical profiles**", "**based spatio**", "**calibration method**", "**motion stereo**", "**spatio temporal**", "**relocation method**", "**stereo images**", "**sensitive noise**", "**shape recovery**", "**wind velocity**", "**method sensitive**" (2) "**c n h**" (4) "**based spatio temporal**", "**d shape recovery**", "**classical spectral element**", "**method sensitive noise**" (2) focuses on numerical methods for constructing images.

(11) Cluster 26) "**p**" (26) "**t**" (16) "**data**", "**set**", "**search**" (13) "**c**", "**model**", "**bar**" (11) "**mass**", "**pb**", "**gev**" (10) "**95**" (9) "**collected**", "**production**", "**e**", "**limits**", "**tev**" (8) "**new**" (7) "**z**", "**l**", "**tevatron**", "**physics**" (6) "**p bar**", "**p p**" (10) "**c l**" (6) "**95 c**", "**roots tev**", "**fermilab tevatron**", "**standard model**" (5) "**data collected**", "**high p**", "**new high**", "**vector leptoquarks**", "**omega t**", "**rho t**", "**bar collisions**", "**channel process**", "**t physics**", "**confidence level**", "**95 confidence**", "**p t**", "**process p**" (4) "**p p bar**" (10) "**95 c l**" (5) "**p bar collisions**", "**p t physics**", "**process p p**", "**high p t**", "**new high p**", "**95 confidence level**", "**channel process p**" (4) "**d empty set**", "**omega t z**", "**t omega t**", "**rho t omega**", "**fermilab tevatron collider**" (3) focuses on new particles from high energy beam collisions, emphasizing research done at the Tevatron.

(11) Cluster 27) "**lead**" (127) "**blood**" (55) "**mug**" (30) "**levels**" (27) "**bone**" (26) "**maternal**" (20) "**dl**" (18) "**exposure**", "**children**", "**pbb**" (16) "**plasma**" (15) "**age**", "**measured**", "**theta**" (12) "**concentration**" (11) "**months**", "**concentrations**" (10) "**level**", "**power**", "**month**" (9) "**risk**", "**increase**", "**mexico**", "**weight**", "**air**", "**school**", "**mean**", "**environmental**", "**tibia**" (8) "**g**", "**m**", "**l**", "**relative**", "**seasonal**", "**questionnaire**", "**pb**", "**fetal**" (7) "**blood lead**" (33) "**lead levels**" (21) "**mug dl**" (18) "**bone lead**" (15) "**lead exposure**", "**lead concentration**" (10) "**plasma lead**" (8) "**theta power**" (7) "**cord blood**", "**whole blood**", "**month age**" (6) "**relative theta**", "**calcium intake**", "**maternal bone**", "**mug l**", "**lead level**" (5) "**blood lead levels**" (11) "**maternal bone lead**", "**relative theta power**" (5) "**blood lead level**", "**blood lead concentration**", "**c academic press**", "**fetal lead exposure**", "**whole blood lead**", "**cord blood lead**" (4)

"**maternal whole blood**", "**lead levels mug**", "**54 72 months**", "**weight month age**", "**maternal plasma lead**", "**atomic absorption spectrophotometry**", "**daily calcium intake**", "**x ray fluorescence**", "**umbilical cord blood**", "**long term memory**", "**population blood lead**" (3) focuses on blood lead level determination, especially in children.

(11) Cluster 28) "**films**" (47) "**thin**" (23) "**optical**" (18) "**bath**" (16) "**deposition**" (15) "**band**", "**gap**", "**cds**", "**cbd**" (12) "**chemical**", "**thickness**" (11) "**field**", "**ev**", "**magnetic**" (10) "**c**" (9) "**energy**", "**e**", "**film**", "**structural**", "**cus**", "**sb2s3**" (8) "**growth**", "**properties**", "**substrates**" (7) "**formation**", "**method**", "**glass**", "**technique**", "**deposited**" (6) "**science**", "**degreesc**", "**application**", "**layer**", "**grown**", "**mum**", "**ternary**", "**ultrasonic**", "**vibration**" (5) "**thin films**" (18) "**band gap**" (12) "**chemical bath**" (10) "**bath deposition**" (8) "**magnetic field**", "**glass substrates**", "**gap energy**" (6) "**ultrasonic vibration**", "**c science**", "**cds thin**" (5) "**deposition cbd**", "**films grown**", "**sb2s3 cus**", "**cbd technique**", "**optical band**", "**films sb2s3**", "**optical properties**" (4) "**films chemical**", "**vibration films**", "**e dir**", "**g e**", "**first stages**", "**b v**", "**science b**", "**zno cds**", "**e g**", "**deposited glass**", "**thin film**", "**substrates chemical**", "**cds films**" (3) "**chemical bath deposition**" (8) "**band gap energy**" (6) "**cds thin films**" (5) "**optical band gap**", "**bath deposition cbd**", "**thin films sb2s3**" (4) "**glass substrates chemical**", "**thin films grown**", "**substrates chemical bath**", "**c science b**", "**ultrasonic vibration films**", "**g e dir**", "**deposited glass substrates**", "**e g e**", "**science b v**" (3) focuses on thin films grown by chemical bath deposition, emphasizing optical properties.

(11) Cluster 29) "**x**" (35) "**c**" (12) "**v**" (11) "**b**", "**science**", "**diffraction**" (10) "**ray**", "**xrd**", "**damage**" (9) "**temperature**", "**ions**" (8) "**chemical**", "**lattice**", "**pl**", "**fwhm**" (7) "**two**", "**growth**", "**crystalline**", "**samples**", "**solid**" (6) "**order**", "**experimental**", "**single**", "**crystal**", "**copper**", "**mev**", "**deuterium**", "**fluences**" (5) "**b v**", "**c science**", "**science b**" (10) "**x ray**" (9) "**ray diffraction**" (7) "**diffraction xrd**", "**solid solution**" (4) "**radiation damage**", "**xrd fwhm**", "**deviations random**", "**fwhm pl**", "**room temperature**", "**light scattering**", "**copper oxides**", "**au ions**", "**x x**" (3) "**science b v**", "**c science b**" (10) "**x ray diffraction**" (7) "**ray diffraction xrd**" (4) "**behaviour hopg diamond**", "**light scattering techniques**", "**sicl4 input ratio**", "**alcl3 sicl4 input**", "**xrd fwhm pl**", "**mev au ions**", "**dynamic light scattering**", "**40 x 40**", "**configurational crystalline order**", "**equal x equal**" (2) focuses on quality and dakage to crystals, emphasizing x-ray diffraction.

(11) Cluster 30) "**insulin**" (43) "**p**" (26) "**resistance**", "**women**" (17) "**men**" (16) "**bmi**", "**glucose**" (15) "**blood**", "**group**", "**fasting**" (14) "**pressure**" (13) "**ratio**", "**index**" (12) "**high**", "**groups**", "**population**", "**levels**", "**concentrations**" (11) "**patients**" (10) "**l**", "**mexican**", "**obese**", "**diabetes**", "**mmol**" (9) "**lower**", "**model**", "**individuals**", "**years**", "**body**", "**equal**", "**cholesterol**", "**pf**" (8) "**insulin resistance**" (17) "**blood pressure**" (13) "**mmol l**" (9) "**mass index**", "**fig ratio**", "**body mass**" (6) "**p 001**", "**pf dm**", "**fasting insulin**", "**type diabetes**" (5) "**insulin sensitivity**", "**index bmi**", "**lp levels**", "**dm eh**", "**p 01**", "**model insulin**", "**early insulin**", "**cell function**", "**fasting glucose**", "**de 27**", "**insulin action**" (4) "**body mass index**" (6) "**pf dm eh**", "**model insulin resistance**", "**mass index bmi**" (4) "**high early insulin**", "**early insulin secretion**", "**function insulin**

**resistance**", "**bmi de 27**", "**cell function insulin**", "**beta cell function**" (3) focuses on clinical studies of insulin resistance, and its relation to a number of chronic diseases.

(11) Cluster 31) "**protein**" (67) "**diet**" (55) "**fed**" (40) "**diets**" (35) "**growth**" (34) "**shrimp**" (27) "**g**" (23) "**higher**" (20) "**l**" (19) "**energy**", "**ratio**" (18) "**weight**" (17) "**low**", "**dietary**", "**high**", "**containing**", "**cbh**" (16) "**levels**" (15) "**p**", "**mg**", "**activity**", "**juveniles**" (14) "**feed**" (13) "**lower**", "**fish**", "**meal**", "**vannamei**", "**salinity**", "**sbm**" (12) "**shrimps**" (11) "**rate**", "**larvae**", "**parts**", "**control**", "**e**", "**carbohydrate**", "**kj**", "**setiferus**" (10) "**l vannamei**" (10) "**p e**", "**weight gain**", "**growth rate**" (9) "**low cbh**", "**l setiferus**", "**parts thousand**", "**mg kj**" (8) "**shrimps fed**", "**g kg**", "**diet containing**", "**control diet**" (7) "**e ratio**", "**soybean soapstock**", "**dietary protein**", "**kg protein**", "**o n**", "**diets containing**", "**artificial diet**", "**thousand salinity**", "**protein levels**" (6) "**p 05**", "**post larvae**", "**fish fed**", "**shrimp fed**", "**36 mg**" (5) "**parts thousand salinity**", "**p e ratio**", "**g kg protein**" (6) "**36 mg kj**" (5) "**400 g kg**", "**dietary protein levels**", "**science b v**", "**extruded canola meal**", "**fed low cbh**", "**c science b**", "**15 parts thousand**", "**shrimps fed low**", "**soluble protein content**", "**40 parts thousand**" (4) focuses on the protein quality of diets fed to small animals, mainly shrimp.

(11) Cluster 32) "**si**" (36) "**surface**" (20) "**atoms**" (17) "**first**", "**calculations**" (14) "**group**" (13) "**two**", "**surfaces**" (12) "**total**", "**110**", "**structure**", "**energy**", "**dimers**", "**001**" (10) "**v**", "**atom**", "**pb**", "**adsorption**" (9) "**c**", "**ga**", "**monolayer**" (8) "**states**", "**single**", "**atomic**", "**principles**", "**ge**" (7) "**b**", "**density**", "**iii**", "**site**", "**coverage**", "**adatom**", "**stm**" (6) "**total energy**", "**si 001**", "**energy calculations**" (8) "**principles total**", "**first principles**" (7) "**group iii**", "**si ge**", "**group v**" (6) "**c 4x8**", "**si dimers**" (5) "**surface states**", "**si 110**", "**110 surfaces**" (4) "**scanning tunneling**", "**monolayer coverage**", "**pb si**", "**d band**", "**top si**", "**phys rev**", "**microscopy stm**", "**iii group**", "**adsorption single**", "**ga atoms**", "**tunneling microscopy**", "**ab initio**" (3) "**first principles total**", "**principles total energy**", "**total energy calculations**" (7) "**tunneling microscopy stm**", "**group iii group**", "**scanning tunneling microscopy**", "**iii group v**" (3) focuses on numerical simulations of adsorption of adatoms on surfaces, mainly silicon, emphasizing first-principles total-energy calculations.

(10) Cluster 33) "**pi**" (73) "**d**" (26) "**xi**" (24) "**k**" (23) "**c**" (17) "**gamma**" (14) "**decays**" (13) "**s**" (11) "**lambda**" (10) "**data**", "**decay**", "**bar**" (9) "**alpha**" (8) "**experiment**", "**fermilab**" (7) "**find**" (6) "**sample**", "**f**", "**sigma**", "**rho**" (5) "**signal**", "**fixed**", "**focus**", "**pk**", "**branching**", "**br**", "**ksk**" (4) "**pi pi**" (30) "**k pi**" (13) "**d k**", "**xi c**" (12) "**gamma d**" (11) "**pi gamma**" (8) "**d s**", "**pi decays**" (6) "**s pi**" (5) "**pk pi**", "**br xi**", "**lambda alpha**", "**alpha xi**", "**k k**", "**alpha lambda**", "**k s**", "**c pk**" (4) "**lambda bar**", "**pi f**", "**sigma 500**", "**b v**", "**ksk pi**", "**fixed target**", "**c science**", "**suppressed decay**", "**xi bar**", "**science b**", "**decays find**", "**xi alpha**" (3) "**pi pi pi**" (8) "**pi pi gamma**" (7) "**d k pi**" (6) "**pi gamma d**", "**s pi pi**" (5) "**gamma d s**", "**c pk pi**", "**k pi pi**", "**lambda alpha xi**", "**gamma d k**", "**xi c pk**", "**d k s**", "**br xi c**", "**pi pi decays**" (4) focuses on radioactive decays, mainly hyperon emphasizing pi-pi production, mainly with the Fermilab database.

(10) Cluster 34) "**group**" (55) "**patients**" (31) "**two**", "**control**" (21) "**h**" (19) "**groups**", "**mg**", "**iu**" (18) "**treated**" (15) "**i**", "**activity**", "**women**" (13) "**l**", "**dose**", "**received**",

"**asa**", "**alcoholic**" (12) "**p**", "**allopurinol**" (11) "**mothers**" (10) "**rats**", "**levels**", "**depigmentation**", "**fsh**" (9) "**low**", "**days**", "**c**", "**years**", "**ewes**", "**nmol**", "**clearance**", "**150**", "**creatinine**" (8) "**control group**" (15) "**creatinine clearance**" (8) "**alcoholic patients**", "**250 iu**", "**l arginine**" (7) "**group i**", "**mode interaction**", "**two groups**", "**style mode**", "**clearance rate**", "**asa activity**", "**experimental group**" (6) "**nmol ml**", "**150 iu**" (5) "**adverse reactions**", "**arylsulfatase asa**", "**ml h**", "**group b**", "**iu l**", "**c science**", "**chronic alcoholic**", "**diabetic rats**", "**mg dl**" (4) "**creatinine clearance rate**", "**style mode interaction**" (6) "**nmol ml h**", "**chronic alcoholic patients**" (4) "**arylsulfatase asa activity**", "**women lower fsh**", "**mg protein h**", "**nmol mg protein**", "**naloxone treated ewes**", "**divided two groups**", "**telangiectasias reticular veins**", "**patients alcohol cirrhosis**", "**obese menopausal women**" (3) focuses on clinical studies of various chronic diseases, emphasizing women and especially alcoholics.

(10) Cluster 35) "**c**" (11) "**american**", "**physics**" (10) "**center**" (7) "**iii**" (6) "**intensity**", "**films**", "**point**", "**lens**" (5) "**density**", "**laser**", "**si**", "**film**", "**deposition**", "**rings**", "**solutions**", "**fluence**", "**ablation**" (4) "**measurements**", "**field**", "**state**", "**electron**", "**light**", "**microscopy**", "**deposited**", "**granules**", "**edge**", "**defect**", "**hillocks**" (3) "**c american**", "**american physics**" (10) "**melting ablation**", "**ophthalmic lens**", "**refractive power**", "**magnetic field**", "**fluence levels**", "**intensity ratios**", "**geo2 sio2**", "**interference rings**", "**ge si**", "**photoluminescence intensity**", "**film deposition**", "**thin film**", "**si site**", "**localized oscillations**", "**line intensity**", "**pyramidal hillocks**", "**antisite defect**" (2) "**c american physics**" (10) "**line intensity ratios**", "**thin film deposition**", "**ge si site**" (2) focuses on a variety of optics problems, emphasizing diagnostics of thin films and monolayers.

(10) Cluster 36) "**cells**" (80) "**t**" (30) "**b**" (23) "**cell**" (21) "**cd4**" (19) "**lymphocytes**", "**anti**" (18) "**cd8**" (17) "**expression**" (14) "**spleen**" (13) "**cd40**" (12) "**mice**", "**il**" (11) "**membranes**", "**cd38**" (10) "**response**" (9) "**apoptosis**", "**proliferation**", "**stimulated**", "**antigen**", "**bcg**", "**cd154**" (8) "**dendritic**", "**levels**", "**dat**", "**expressed**", "**dmf**", "**cd40l**" (7) "**m**", "**dependent**", "**vitro**", "**activated**", "**vivo**", "**human**", "**mouse**", "**langerhans**", "**f2**", "**b220**", "**igm**" (6) "**t cells**" (20) "**cd4 cd8**", "**b cells**" (11) "**spleen cells**" (7) "**langerhans cells**", "**t cell**" (6) "**dendritic cells**", "**mouse spleen**", "**cd8 lymphocytes**" (5) "**f2 dmf**", "**b cell**", "**b lymphocytes**", "**l929 cells**", "**cd8 t**", "**cd4 t**", "**class ii**", "**cd40l expression**", "**mycoplasma membranes**" (4) "**cd4 cd8 lymphocytes**" (5) "**mouse spleen cells**", "**cd8 t cells**", "**cd4 t cells**" (4) "**class ii molecules**", "**major histocompatibility complex**", "**apoptosis cd4 t**", "**mammalian langerhans cells**" (3) focuses on studying cells that affect immune system response.

(10) Cluster 37) "**degreesc**", "**corrosion**" (28) "**temperature**" (12) "**c**", "**s**", "**steel**" (11) "**electrochemical**", "**material**" (10) "**x**" (9) "**m**", "**temperatures**", "**electron**", "**900**", "**fracture**", "**resistance**", "**700**" (8) "**hot**", "**rate**", "**rates**", "**carried**", "**wear**" (7) "**two**", "**v**", "**surface**", "**time**", "**techniques**", "**heat**", "**tests**", "**scanning**" (6) "**50**", "**process**", "**decrease**", "**ray**", "**mechanism**", "**higher**", "**treated**", "**intergranular**", "**microscopy**", "**velocities**", "**sem**" (5) "**900 degreesc**", "**700 degreesc**" (7) "**scanning electron**", "**corrosion rate**" (6) "**x ray**", "**electron microscopy**" (5) "**carbon steel**", "**hot ductility**", "**velocities m**", "**c science**", "**m s**" (4) "**fracture toughness**", "**surface states**", "**electron**

**microscope**", "**strain rates**", "**corrosion resistance**", "**hot corrosion**", "**corrosion process**", "**heat treatment**", "**x x**", "**degreesc 900**", "**625 700**", "**weight loss**", "**fe40al 1b**", "**aging time**", "**v vs**", "**heat treated**", "**tests carried**", "**treated material**" (3) "**velocities m s**" (4) "**degreesc 900 degreesc**", "**scanning electron microscopy**", "**625 700 degreesc**", "**scanning electron microscope**" (3) focuses on corrosion behavior of steel.

(10) Cluster 38) "**population**" (44) "**density**" (34) "**growth**" (33) "**food**" (32) "**x**" (29) "**ml**", "**chlorella**" (27) "**rate**" (22) "**cells**" (20) "**densities**" (18) "**l**" (16) "**d**" (15) "**b**" (14) "**water**", "**concentrations**" (13) "**day**", "**increase**", "**patulus**" (12) "**concentration**", "**rotifers**", "**microcystis**" (11) "**algal**" (10) "**higher**", "**mg**", "**shrimp**", "**abundance**" (9) "**levels**", "**exchange**", "**fed**", "**reproductive**" (8) "**two**", "**lower**", "**survival**", "**highest**", "**varied**", "**r**", "**ind**" (7) "**population growth**" (25) "**cells ml**" (18) "**x cells**" (12) "**b patulus**", "**x x**" (10) "**rate population**" (9) "**food density**", "**water exchange**" (8) "**ind ml**", "**mg l**", "**population increase**" (7) "**chlorella vulgaris**", "**reproductive rate**", "**growth rate**" (6) "**methyl parathion**", "**d pulex**", "**increase day**", "**population densities**", "**food levels**" (5) "**x cells ml**" (12) "**rate population increase**", "**x x cells**" (7) "**population increase day**" (5) "**green alga chlorella**", "**alga chlorella vulgaris**", "**population growth b**", "**net reproductive rate**" (4) "**spotted sand bass**", "**growth b patulus**", "**density water exchange**", "**population growth rotifer**", "**increase day r**" (3) focuses on population growth of small animals in aquatic environments, related to mainly algae concentrations.

(10) Cluster 39) "**c**" (24) "**n**" (22) "**nmr**" (21) "**h**" (19) "**p**" (18) "**31**" (16) "**15**" (14) "**13**", "**chemical**", "**se**" (9) "**coupling**", "**constants**", "**sons**" (8) "**s**", "**copyright**", "**shifts**" (7) "**degrees**", "**j**", "**77**", "**sign**", "**ring**", "**nh**", "**imino**" (6) "**b**", "**14**", "**delta**", "**o**" (5) "**structures**", "**solution**", "**experiments**", "**parameters**", "**compounds**", "**sulfur**", "**conformation**", "**pyridine**", "**phosphorus**" (4) "**p 31**" (16) "**c 13**" (9) "**n 15**", "**c sons**" (8) "**coupling constants**", "**15 p**", "**copyright c**", "**chemical shifts**" (7) "**h c**", "**se 77**" (6) "**n 14**", "**14 15**", "**delta n**" (5) "**nh pyridine**", "**h nmr**", "**13 nmr**" (4) "**31 se**", "**shifts delta**", "**77 p**", "**13 n**", "**isotope induced**", "**nmr parameters**", "**nmr spectroscopy**", "**sulfur selenium**", "**x ray**", "**induced chemical**", "**membered ring**", "**si 29**", "**77 nmr**" (3) "**15 p 31**", "**copyright c sons**" (7) "**delta n 14**", "**h c 13**", "**n 14 15**" (5) "**c 13 nmr**", "**14 15 p**" (4) "**c 13 n**", "**se 77 p**", "**induced chemical shifts**", "**isotope induced chemical**", "**13 n 15**", "**shifts delta n**", "**31 se 77**", "**p 31 se**", "**77 p 31**", "**se 77 nmr**", "**chemical shifts delta**", "**n 15 p**" (3) focuses on characteriziation of molecular structures of compounds, mainly phosphanes using mainly NMR.

(10) Cluster 40) "**group**" (35) "**p**" (25) "**treated**", "**smb**" (20) "**groups**", "**cows**" (19) "**05**", "**salmonella**" (15) "**control**" (13) "**days**", "**c**", "**women**", "**enteritidis**" (11) "**day**", "**h**", "**acid**", "**serum**", "**uric**" (10) "**first**", "**concentrations**", "**animals**", "**obese**", "**hepatica**" (9) "**s**", "**treatment**", "**difference**", "**progesterone**" (8) "**f**", "**lower**", "**rates**", "**mg**", "**rbst**", "**file**", "**flaring**" (7) "**p 05**" (15) "**salmonella enteritidis**" (11) "**control group**", "**uric acid**" (10) "**obese women**", "**f hepatica**" (7) "**c science**" (6) "**smb ccr**", "**progesterone concentrations**", "**difference p**" (5) "**bursa fabricius**", "**mg kg**", "**e s**" (4) "**difference p 05**" (4) "**women control group**", "**non obese women**", "**obese women**

**control**", "**tubular secretion uric**", "**groups p 05**", "**secretion uric acid**", "**file fit apex**", "**cows smb ccr**", "**c science b**", "**science b v**", "**e s antigens**" (3) focuses both on diagnosis and treatment of infections in animals and humans as well as examination of embryo growth factors.

(10) Cluster 41) "**degreesc**" (27) "**samples**" (16) "**electron**" (11) "**x**", "**ray**", "**diffraction**" (10) "**microscopy**", "**zirconia**" (9) "**c**", "**nm**", "**scanning**", "**powder**" (8) "**surface**", "**ldpe**" (7) "**ethanol**", "**method**", "**process**", "**sol**", "**gel**", "**thermal**", "**formed**", "**crystallization**", "**tetragonal**" (6) "**500**", "**increased**", "**science**", "**carbon**", "**area**", "**high**", "**phase**", "**spectroscopy**", "**crystalline**", "**transformation**", "**diameter**", "**amorphous**", "**sem**", "**nanotubes**", "**mullite**" (5) "**x ray**" (10) "**electron microscopy**" (9) "**scanning electron**", "**ray powder**", "**sol gel**" (6) "**microscopy sem**", "**powder diffraction**", "**carbon nanotubes**" (5) "**surface area**", "**c science**", "**800 degreesc**", "**ray diffraction**" (4) "**samples x**", "**500 degreesc**", "**gave rise**", "**morphological structural**", "**b v**", "**tetragonal monoclinic**", "**science b**", "**samples ethanol**", "**tetragonal zirconia**", "**infrared spectroscopy**" (3) "**x ray powder**", "**scanning electron microscopy**" (6) "**electron microscopy sem**", "**ray powder diffraction**" (5) "**x ray diffraction**" (4) "**samples x ray**", "**science b v**", "**c science b**" (3) "**xrd fourier transform**", "**oxides calcined 500**", "**degreesc sulphate ion**", "**transmission electron microscopy**", "**ray diffraction xrd**", "**biologically treated ldpe**", "**100 1400 degreesc**", "**dta tga scanning**", "**formed nanometric particles**", "**calcined 500 degreesc**", "**average crystallite sizes**" (2) focuses on study of material properties and surface conditions using x-ray powder diffraction and scanning electron microscopy.

(10) Cluster 42) "**electron**" (25) "**energy**" (22) "**cross**" (20) "**sections**" (14) "**experimental**" (13) "**single**", "**loss**" (12) "**capture**" (10) "**agreement**" (9) "**data**", "**kev**" (8) "**range**", "**total**", "**n**" (7) "**good**", "**energies**", "**theory**", "**section**" (6) "**order**", "**measured**", "**mass**", "**atomic**", "**spheres**" (5) "**measurements**", "**interaction**", "**r**", "**width**", "**electrons**" (4) "**cross sections**" (14) "**single electron**" (11) "**electron loss**" (8) "**energy range**", "**good agreement**", "**cross section**", "**electron capture**" (6) "**loss cross**" (5) "**sections single**", "**electron energy**", "**energy loss**", "**total cross**", "**capture cross**", "**range kev**" (4) "**agreement experimental**", "**experimental data**" (3) "**single electron loss**" (6) "**single electron capture**", "**electron loss cross**" (5) "**electron energy loss**", "**cross sections single**", "**energy range kev**", "**loss cross sections**", "**sections single electron**", "**total cross sections**" (4) "**good agreement experimental**" (3) focuses on determination of electron cross sections, emphasizing single electron capture and single electron loss.

(10) Cluster 43) "**field**", "**magnetic**" (37) "**b**" (24) "**potential**" (12) "**electron**" (11) "**magnetoresistance**" (10) "**h**", "**x**", "**dependence**", "**linear**" (8) "**two**", "**interactions**", "**stationary**", "**pm**" (7) "**wave**", "**approximate**", "**contacts**" (6) "**alpha**", "**conditions**", "**method**", "**temperature**", "**equation**", "**fluctuations**", "**mean**", "**afm**" (5) "**magnetic field**" (26) "**field b**" (5) "**vector potential**", "**x b**", "**pm afm**", "**b approximate**", "**bounded semiconductors**", "**spatial dependence**" (4) "**quadrupole interactions**", "**electric potential**", "**del x**", "**mean magnetic**", "**electron fluid**", "**protoplanetary disks**", "**linear dependence**", "**dependence potential**", "**potential contacts**",

"**distribution electric**" (3) "**magnetic field b**" (5) "**del x b**", "**mean magnetic field**", "**distribution electric potential**", "**spatial dependence potential**" (3) "**external magnetic field**", "**2d electron fluid**", "**electric potential contacts**", "**magnetoresistance bounded semiconductors**", "**b del x**", "**pm afm phase**", "**mean free path**", "**fm pm afm**", "**b approximate b**", "**semiconductors linear contribution**", "**electron temperature distribution**", "**bounded semiconductors linear**", "**dependence magnetic field**", "**b r t**" (2) focuses on motions of ions and electrons in fluctuating magnetic fields.

(10) Cluster 44) **h**" (27) "**ii**" (25) "**hh**" (22) "**velocity**" (20) "**regions**" (16) "**jet**" (15) "**nebula**" (13) "**emission**", "**similar**" (12) "**observations**", "**stars**" (11) "**line**", "**high**", "**lines**", "**bipolar**" (9) "**alpha**", "**star**" (8) "**density**", "**central**", "**n**", "**s**", "**model**", "**wind**", "**ionizing**", "**bubbles**" (7) "**stellar**", "**across**", "**features**", "**axis**", "**hydrogen**", "**204**", "**shock**", "**outflows**" (6) "**h ii**" (18) "**ii regions**" (14) "**n ii**", "**h alpha**" (7) "**km s**" (5) "**ii region**", "**radial velocity**", "**hh 202**", "**central star**" (4) "**similar equal**", "**bow shock**", "**alpha n**", "**o stars**", "**deuterium hydrogen**", "**collimated outflows**", "**n11b n180b**", "**203 204**", "**hh 204**", "**blown bubbles**", "**electron density**", "**wind blown**", "**hh 111**", "**planetary nebulae**", "**extended emission**", "**jetlike features**" (3) "**h ii regions**" (14) "**h ii region**" (4) "**h alpha n**", "**wind blown bubbles**", "**alpha n ii**" (3) "**hh 203 204**", "**high dispersion echelle**", "**extended emission radial**", "**main sequence o**", "**symmetry bipolar lobes**", "**emission radial velocity**", "**hh 202 part**", "**sequence o stars**", "**systemic velocity bipolar**", "**bubbles main sequence**", "**hubble space telescope**", "**269 features hh**", "**bubbles n11b n180b**", "**ii regions bubbles**" (2) focuses on observation of phenomena in stars and nebulae, mainly by spectroscopic instruments.

(10) Cluster 45) "**hpv**" (57) "**women**", "**cervical**" (26) "**cancer**" (20) "**infection**" (19) "**patients**", "**group**", "**dna**" (18) "**p**" (17) "**risk**", "**levels**" (15) "**age**", "**types**" (14) "**control**", "**mexico**", "**years**", "**variants**" (13) "**high**", "**prevalence**", "**95**", "**e6**", "**aav**" (12) "**cases**", "**e**", "**case**", "**proteins**", "**papillomavirus**", "**e7**", "**foot**" (11) "**aa**", "**e2**" (10) "**low**", "**invasive**", "**damage**", "**human**", "**serum**", "**factors**", "**ulcers**" (9) "**cervical cancer**" (17) "**hpv infection**", "**hpv types**" (12) "**e6 e7**" (10) "**dna damage**" (9) "**foot ulcers**" (8) "**odds ratio**", "**human papillomavirus**", "**e7 proteins**" (7) "**invasive cervical**", "**serum magnesium**", "**e2 e6**", "**aa variants**", "**papillomavirus hpv**" (6) "**united states**", "**case patients**", "**e variants**", "**hpv dna**" (5) "**e6 e7 proteins**" (7) "**e2 e6 e7**", "**invasive cervical cancer**", "**human papillomavirus hpv**" (6) "**95 confidence interval**" (4) "**polymerase chain reaction**", "**cervical cancer mexico**", "**united states mexico**", "**type diabetes foot**", "**e7 proteins patients**", "**risk hpv types**", "**diabetes foot ulcers**" (3) focuses on correlation of viruses and genes with various abnormalities, mainly cervical cancer.

(10) Cluster 46) "**c**" (19) "**electron**" (15) "**energy**" (13) "**temperature**" (12) "**carbon**" (11) "**b**", "**n**", "**spectroscopy**" (9) "**science**", "**k**", "**1s**" (8) "**v**", "**t**", "**band**" (7) "**pi**", "**x**", "**loss**", "**core**", "**films**" (6) "**spectra**", "**chemical**", "**structure**", "**nm**", "**peaks**" (5) "**c science**" (8) "**b v**", "**science b**" (7) "**energy loss**", "**electron energy**" (5) "**room temperature**", "**x ray**", "**n 1s**", "**1s c**", "**c 1s**" (4) "**loss spectroscopy**", "**1073 k**", "**core shell**", "**hybridized carbon**" (3) "**science b v**", "**c science b**" (7) "**electron energy loss**" (5) "**n 1s c**", "**1s c 1s**" (4) "**energy loss spectroscopy**" (3) "**c american physics**", "**c 1s**

**regions**", "**core level photoemission**", "**sp hybridized carbon**", "**x ray photoelectron**", "**823 1073 k**" (2) focuses on characterization of carbon species, emphasizing use of electron energy loss spectroscopy.

(10) Cluster 47) "**control**" (34) "**pid**" (12) "**stability**" (10) "**pi**", "**paper**" (9) "**c**", "**classical**" (7) "**science**", "**based**", "**chemical**" (6) "**two**", "**dynamics**", "**experimental**", "**loop**", "**global**", "**pd**", "**stabilization**", "**compensation**" (5) "**size**", "**direct**", "**proportional**", "**regulation**", "**scheme**", "**integral**", "**robot**", "**torque**", "**drive**", "**controllers**", "**reactors**" (4) "**pid control**" (8) "**c science**" (6) "**pi control**", "**pd control**" (5) "**control scheme**", "**classical pi**", "**direct drive**", "**stabilization chemical**", "**chemical reactors**" (4) "**science b**", "**control configuration**", "**b v**", "**proportional integral**" (3) "**science b v**", "**c science b**" (3) focuses on feedback control of systems, primarily chemical reactors and robotics.

(9) Cluster 48) "**ca2**" (53) "**type**" (26) "**channel**" (23) "**channels**" (21) "**l**" (17) "**inhibition**" (15) "**g**", "**current**", "**currents**", "**fibers**" (11) "**slow**", "**muscle**" (10) "**cytosolic**" (9) "**i**", "**k**", "**voltage**", "**skeletal**", "**sv**" (8) "**non**", "**protein**", "**muscarinic**", "**denervated**" (6) "**two**", "**modulation**", "**n**", "**dependent**", "**cell**", "**properties**", "**amplitude**", "**anions**", "**pma**", "**inactivation**", "**depolarization**", "**na**", "**component**", "**rgs2**" (5) "**type ca2**" (18) "**l type**" (17) "**ca2 channels**" (11) "**ca2 channel**" (9) "**ca2 currents**", "**skeletal muscle**" (8) "**n type**", "**cytosolic ca2**", "**denervated fibers**" (5) "**component inhibition**", "**muscarinic inhibition**", "**muscle fibers**", "**alterations current**", "**non denervated**", "**sv channel**", "**slow muscarinic**", "**g protein**", "**type channels**" (4) "**l type ca2**" (14) "**type ca2 channels**" (10) "**n type ca2**", "**type ca2 currents**", "**skeletal muscle fibers**" (4) "**non denervated fibers**", "**l type channels**", "**slow muscarinic inhibition**", "**long term depolarization**", "**alterations current kinetics**", "**type ca2 channel**" (3) focuses on manipulation of clacium currents in channels.

(9) Cluster 49) "**emission**", "**star**" (18) "**similar**" (17) "**source**", "**region**", "**maser**" (14) "**observations**" (12) "**radio**" (11) "**s**", "**outflow**" (10) "**sources**", "**h2o**" (9) "**located**", "**shock**", "**au**", "**equal**", "**binary**", "**protostar**" (8) "**stars**", "**m**", "**system**", "**mass**", "**south**", "**mas**", "**sio**", "**hw2**" (7) "**formation**", "**morphology**", "**j**", "**mum**", "**iras**", "**masers**", "**g138**" (6) "**similar equal**" (8) "**maser emission**" (6) "**g138 295**", "**295 555**" (5) "**bipolar outflow**", "**iras 16293**", "**16293 2422**", "**star formation**", "**h2o maser**" (4) "**ob type**", "**1915 105**", "**555 s**", "**binary system**", "**large array**", "**hd 97950**", "**rs vir**", "**grs 1915**", "**type stars**", "**maser variations**", "**afgl 4029**", "**km s**" (3) "**g138 295 555**" (5) "**iras 16293 2422**" (4) "**ob type stars**", "**295 555 s**", "**grs 1915 105**" (3) focuses on emissions from starts, especially in the maser spectrum.

(9) Cluster 50) "**t**" (75) "**x**" (19) "**n**" (18) "**systems**" (14) "**alpha**" (13) "**time**" (12) "**g**", "**two**", "**c**" (11) "**hr**" (9) "**system**", "**quantum**" (8) "**w**" (7) "**model**", "**length**", "**function**", "**flow**", "**states**" (6) "**f**", "**s**", "**delta**", "**particle**", "**case**", "**gamma**", "**particles**", "**random**", "**infinity**", "**increases**", "**tau**", "**gg**" (5) "**t t**" (15) "**x t**" (9) "**w t**" (6) "**t n**", "**n t**", "**n hr**" (5) "**infinity x**", "**flow patterns**", "**t coll**", "**s t**", "**n x**" (4) "**t similar**", "**t f**", "**t g**", "**fluctuation limits**", "**g t**", "**t xi**", "**y t**", "**t y**", "**t eta**", "**eta t**", "**c**

**hierarchical**", "**alpha x**", "**xi t**", "**occupation time**", "**random walk**", "**hierarchical random**" (3) "**t n hr**", "**n t n**" (5) "**t xi t**", "**x t t**", "**x t f**", "**y t g**", "**hierarchical random walk**", "**t t xi**", "**t y t**", "**xi t y**", "**x t eta**", "**t eta t**", "**c hierarchical random**" (3) focuses on excited many particle states operating under random conditions.

(9) Cluster 51) "**cells**" (40) "**cell**" (23) "**ngf**" (21) "**insulin**" (12) "**secretion**" (10) "**beta**", "**cd95**" (9) "**glucose**", "**pancreatic**", "**cancer**" (8) "**increase**", "**primary**", "**breast**" (7) "**two**", "**growth**", "**expression**", "**l**", "**role**", "**carcinoma**", "**cervical**", "**lrp**" (6) "**c**", "**factor**", "**tryptase**" (5) "**formation**", "**p**", "**receptors**", "**lines**", "**root**", "**factors**", "**mmol**", "**malignant**", "**dbcamp**", "**sf**", "**kit**", "**pericycle**", "**cxcr4**" (4) "**insulin secretion**" (10) "**beta cells**" (7) "**breast cancer**" (6) "**cervical carcinoma**" (5) "**mmol l**", "**pericycle cells**", "**l glucose**" (4) "**pancreatic beta**", "**c kit**", "**cancer cells**", "**pancreatic p**", "**cd95 r**", "**cxcr4 ccr7**", "**cell lines**", "**carcinoma cells**", "**cell interaction**", "**ngf insulin**", "**p cells**" (3) "**mmol l glucose**" (4) "**breast cancer cells**", "**cervical carcinoma cells**", "**ngf insulin secretion**" (3) focuses on cell physiology, especially insulin secretion in pancreatic cells and cervical cancer cell signals.

(9) Cluster 52) "**activity**" (26) "**enzyme**" (25) "**kda**", "**purified**" (10) "**molecular**" (9) "**d**", "**ph**" (8) "**m**", "**degreesc**", "**glucose**", "**lipase**" (7) "**c**", "**l**", "**acid**", "**ildh**" (6) "**k**", "**70**", "**higher**", "**weight**", "**subunits**", "**lactate**" (5) "**two**", "**fold**", "**mm**", "**high**", "**kinase**", "**specific**", "**mass**", "**phosphate**", "**partially**", "**substrate**", "**phosphorylation**", "**tnp**", "**homogeneity**", "**acetonitrile**" (4) "**molecular weight**", "**k m**", "**enzyme activity**" (5) "**molecular mass**" (4) "**lipase activity**", "**phytase activity**", "**c academic**", "**d lactate**", "**purified enzyme**", "**glucose kinase**", "**l d**", "**academic press**", "**70 kda**" (3) "**l d lactate**", "**c academic press**" (3) "**abts k m**", "**dolichol phosphate mannose**", "**var caesius glucose**", "**lactic acid bacteria**", "**peucetius var caesius**", "**caesius glucose kinase**", "**octyl sepharose cl**", "**sepharose cl 4b**" (2) focuses on enzymatic compounds extracted from plants.

(9) Cluster 53) "**induced**" (22) "**liver**" (20) "**administration**" (16) "**ethanol**" (15) "**h**" (14) "**serum**" (13) "**rats**", "**lipid**", "**mpp**" (12) "**increased**", "**peroxidation**", "**anit**" (11) "**activity**", "**cholestasis**" (10) "**levels**", "**acute**", "**hepatic**" (9) "**24**", "**damage**", "**ccl4**", "**cuso4**" (8) "**increase**", "**mg**", "**activities**", "**gdcl3**" (7) "**ph**", "**zn**", "**mice**", "**injected**", "**copper**", "**injury**", "**melatonin**", "**antioxidant**", "**inos**", "**hydroxymelatonin**" (6) "**lipid peroxidation**" (11) "**24 h**" (6) "**anit induced**", "**liver injury**", "**superoxide dismutase**", "**induced liver**", "**mg kg**", "**n acetylserotonin**" (5) "**anit injected**", "**injected rats**", "**melatonin hydroxymelatonin**", "**mpp induced**", "**c science**", "**ethanol administration**" (4) "**hydroxymelatonin n**", "**total bilirubin**", "**serum levels**", "**gsh gssg**", "**kg body**", "**body weight**", "**serum total**", "**neutrophil infiltration**", "**kupffer cells**", "**bilirubin concentration**", "**free radicals**", "**acute administration**" (3) "**anit injected rats**", "**induced liver injury**" (4) "**mg kg body**", "**hydroxymelatonin n acetylserotonin**", "**anit induced liver**", "**serum total bilirubin**", "**total bilirubin concentration**", "**kg body weight**" (3) focuses on determination of liver damage mechanisms from induced toxicity.

(9) Cluster 54) "**pulse**" (18) "**b**", "**c**", "**pulses**" (9) "**v**", "**science**", "**temporal**", "**optical**" (7) "**time**", "**harmonic**" (6) "**frequency**", "**generation**", "**laser**", "**soliton**" (5) "**two**", "**high**", "**raman**", "**fiber**", "**polarization**", "**pump**", "**train**" (4) "**experiment**", "**state**", "**crystal**", "**self**", "**radiation**", "**femtosecond**", "**variations**", "**neutron**", "**multiplicity**" (3) "**c science**", "**science b**", "**b v**" (7) "**harmonic pulse**" (4) "**raman self**", "**high multiplicity**" (3) "**pulse generation**", "**doubling crystal**", "**phys b**", "**function delay**", "**b proc**", "**pump pulse**", "**laser pulses**", "**pulses pump**", "**dual frequency**", "**time fluorescence**", "**suppl 75a**", "**proc suppl**", "**nucl phys**", "**75a 333**", "**self scattering**", "**delay time**", "**plastic scintillators**", "**shortening harmonic**" (2) "**c science b**", "**science b v**" (7) "**nucl phys b**", "**raman self scattering**", "**b proc suppl**", "**shortening harmonic pulse**", "**suppl 75a 333**", "**proc suppl 75a**", "**function delay time**", "**phys b proc**" (2) focuses on the propagation and detection of pulses, mainly from optical sources.

(9) Cluster 55) "**films**" (51) "**sn**" (17) "**x**" (15) "**ray**" (14) "**temperature**", "**pressure**" (9) "**f**", "**growth**", "**o**", "**deposited**", "**sputtering**" (8) "**c**", "**degreesc**", "**diffraction**", "**precursor**", "**fluorine**", "**sno**" (7) "**oxygen**", "**cu**", "**deposition**", "**molar**", "**tyrosinase**", "**vapors**", "**licoo2**", "**sn4**" (6) "**measurements**", "**i**", "**spectroscopy**", "**e**", "**silicon**", "**fraction**", "**substrates**", "**grown**", "**target**", "**plasma**", "**die**", "**superconducting**", "**chitosan**", "**cresol**" (5) "**x ray**" (14) "**sn sn**" (7) "**sn molar**", "**ray diffraction**" (6) "**i e**", "**chitosan films**" (5) "**tyrosinase coated**", "**sn4 sn**", "**o f**", "**films deposited**", "**silicon oxynitride**", "**precursor films**", "**coated chitosan**", "**films grown**", "**oxynitride films**", "**molar fraction**" (4) "**x ray diffraction**" (6) "**sn molar fraction**", "**silicon oxynitride films**", "**sn4 sn molar**", "**tyrosinase coated chitosan**", "**sn sn sn**", "**coated chitosan films**" (4) "**sn sn4 sn**", "**x ray diffractograms**", "**c science b**", "**x ray photoelectron**", "**ba cu o**", "**snox f films**", "**ray photoelectron spectroscopy**", "**science b v**" (3) focuses on deposition and growth of thin films at different pressures and temperatures, emphasizing the use of x-ray variants as diagnostics.

(9) Cluster 56) "**t**" (27) "**mice**" (22) "**vaccine**" (15) "**response**", "**solium**" (14) "**antibodies**", "**dna**" (13) "**immune**", "**protein**", "**plasmid**", "**taenia**", "**protective**" (10) "**proteins**", "**immunization**", "**rop2**" (9) "**cells**", "**cysticercosis**", "**crassiceps**", "**vaccination**", "**shock**", "**immunized**" (8) "**c**", "**kda**", "**heat**", "**recombinant**", "**paramyosin**" (7) "**strain**", "**induced**", "**cell**", "**igg**", "**amino**", "**challenge**", "**fmdv**" (6) "**t solium**" (9) "**immune response**", "**heat shock**" (7) "**shock proteins**", "**t cells**", "**taenia crassiceps**", "**taenia solium**" (5) "**ts strain**", "**t cell**", "**i s**", "**balb c**", "**cellular immune**", "**dna vaccine**" (4) "**c mice**", "**solium paramyosin**", "**cba j**", "**terminal fragment**", "**plasmid dna**", "**strain t**", "**mice immunized**", "**solium cysticercosis**", "**v h**", "**vaccine candidate**", "**s dna**", "**igg 2a**", "**t crassiceps**", "**cell epitope**" (3) "**heat shock proteins**" (5) "**i s dna**", "**balb c mice**" (3) focuses on the development of vaccines for both animals and humans to immunize against a wide variety of viruses.

(9) Cluster 57) "**mrna**" (32) "**insulin**" (26) "**levels**", "**trh**" (25) "**expression**" (23) "**glucokinase**" (21) "**liver**", "**activity**" (17) "**rat**", "**fetal**" (16) "**gene**" (15) "**bcat**" (12) "**day**", "**pancreatic**" (10) "**cells**", "**acid**", "**protein**", "**serum**", "**hepatic**" (9) "**rats**", "**h**", "**testosterone**", "**camp**", "**retinoic**" (8) "**increased**", "**days**", "**adult**", "**vitro**", "**cell**", "**neurons**", "**trkb**", "**bdnf**", "**biotin**", "**bcatm**" (7) "**mrna levels**" (15) "**trh mrna**", "**gene**

**expression**" (12) "**bcat activity**", "**insulin gene**", "**retinoic acid**" (8) "**hepatic glucokinase**", "**rat liver**" (7) "**pro trh**" (6) "**estrous cycle**", "**glucokinase activity**", "**fetal liver**" (5) "**pancreatic glucokinase**", "**serum insulin**", "**insulin mrna**", "**liver nuclei**", "**h incubation**", "**mammary gland**" (4) "**deficient rats**", "**c science**", "**trans retinoic**", "**glucokinase gene**", "**glucose metabolism**", "**kda protein**", "**24 h**", "**trh neurons**", "**rt pcr**", "**expression estrous**" (3) "**trh mrna levels**" (9) "**insulin gene expression**" (8) "**pro trh mrna**" (5) "**expression estrous cycle**", "**trans retinoic acid**", "**gene expression estrous**" (3) focuses on peptide biosynthesis emphasizing TRH mRNA concentrations, as well as insulin mRNA concentrations for analyzing insulin gene expression patterns.

(9) Cluster 58) "**c**" (38) "**h**" (21) "**o**" (16) "**mol**" (15) "**kcal**" (14) "**n**" (10) "**energy**" (9) "**bond**", "**mp2**" (8) "**cc**", "**structure**", "**bonds**", "**protonation**", "**aug**" (7) "**interaction**" (6) "**cytochrome**", "**activation**", "**level**", "**basis**", "**theory**", "**equatorial**", "**hydrogen**", "**dimers**", "**ax**", "**adamantonium**" (5) "**structures**", "**stable**", "**good**", "**phase**", "**set**", "**y**", "**agreement**", "**optimized**", "**point**", "**axial**", "**pvtz**" (4) "**kcal mol**" (14) "**c h**" (8) "**aug cc**" (7) "**c c**" (6) "**mp2 aug**" (5) "**cc pvtz**", "**o c**", "**c o**", "**good agreement**", "**h o**", "**basis set**" (4) "**h c**", "**c bond**", "**energy surfaces**", "**initio electronic**", "**c sp**", "**potential energy**", "**electronic structure**", "**protonation c**", "**optimized geometries**", "**structure theory**", "**sp c**", "**dihedral angle**", "**n methylacetamide**", "**mol stable**" (3) "**mp2 aug cc**" (5) "**aug cc pvtz**" (4) "**electronic structure theory**", "**c sp c**", "**h o c**", "**potential energy surfaces**", "**initio electronic structure**", "**kcal mol stable**" (3) focuses on molecular and electronic structures and energetics of compounds, using a combination of experiemnt and theory.

(9) Cluster 59) "**degreesc**", "**temperature**" (14) "**x**" (13) "**high**", "**implanted**", "**powders**" (12) "**ions**", "**samples**" (11) "**spectra**", "**o**", "**temperatures**", "**optical**", "**metal**" (10) "**solutions**" (9) "**process**", "**precursors**" (8) "**b**", "**emission**", "**c**", "**method**", "**concentration**", "**reaction**", "**ag**", "**nm**", "**purity**", "**eu**" (7) "**absorption**", "**acid**", "**no3**", "**photoluminescence**", "**cu**", "**combustion**", "**silica**" (6) "**high purity**" (7) "**optical properties**", "**x o**" (5) "**low temperature**", "**c science**", "**versatic acid**", "**200 degreesc**", "**implanted samples**" (4) focuses on both the optical and electrical properties of materials with ion implants as well as low temperature methods to produce powders with high purity, high chemical homogeneity and improved crystallinity.

(9) Cluster 60) "**species**" (43) "**total**", "**collected**" (15) "**m**" (13) "**abundance**" (12) "**larvae**" (11) "**samples**", "**gulf**" (10) "**s**", "**mexico**" (9) "**20**", "**colima**" (8) "**individuals**", "**sp**", "**pacific**", "**faw**" (7) "**number**", "**abundant**", "**jalisco**", "**families**", "**michoacan**", "**soil**" (6) "**three**", "**area**", "**states**" (5) "**faw larvae**" (6) "**soil samples**" (5) "**gulf california**" (4) "**lower intertidal**", "**continental shelf**", "**chelonus sp**", "**rate parasitism**", "**relative abundance**", "**cabo pulmo**", "**pacific ocean**", "**s parva**", "**s hancocki**", "**colima jalisco**", "**samples collected**", "**1000 m**", "**number species**", "**intertidal zone**" (3) focuses on the abundance of various fish and organism species, emphasizing those contained in the Gulf of Mexico.

(9) Cluster 61) "**species**" (65) "**diversity**" (14) "**f**", "**richness**" (12) "**forest**" (11) "**season**" (10) "**habitats**" (9) "**increased**", "**mexico**", "**rodent**" (8) "**three**", "**similar**", "**habitat**" (7) "**m**", "**high**", "**population**", "**patterns**", "**sand**", "**endemic**", "**abundance**", "**seed**", "**transects**", "**pronghorn**" (6) "**strong**", "**time**", "**areas**", "**individuals**", "**distribution**", "**dry**", "**vegetation**", "**basin**", "**community**", "**sierra**", "**dispersal**", "**dispersed**", "**mazateca**" (5) "**species richness**" (12) "**dry season**", "**sand movement**", "**diversity increased**" (4) "**pine oak**", "**increased increasing**", "**season dry**", "**bat species**", "**habitat types**", "**rodent species**", "**relative abundance**", "**pronghorn diet**", "**sierra mazateca**", "**number individuals**", "**wet season**" (3) "**wet season dry**", "**season dry season**" (3) focuses on abundance and diversity of species as a function of habitat characteristics.

(9) Cluster 62) "**angstrom**" (28) "**structure**" (19) "**c**" (14) "**x**" (13) "**group**" (11) "**b**", "**space**", "**r**" (10) "**compound**", "**ray**", "**crystal**", "**diffraction**" (8) "**z**", "**degrees**" (7) "**12**", "**atoms**", "**parameters**" (6) "**two**", "**fe**", "**n**", "**beta**", "**sites**" (5) "**g**", "**v**", "**delta**", "**cell**", "**p**", "**o**", "**monoclinic**", "**cm**", "**layers**", "**subsystem**", "**chains**", "**octahedral**", "**octahedra**" (4) "**space group**" (9) "**x ray**" (8) "**angstrom c**" (6) "**angstrom b**" (5) "**ray crystallography**", "**g cm**", "**angstrom beta**", "**r barm**", "**octahedral sites**", "**angstrom z**", "**delta doped**", "**barm space**", "**ray diffraction**" (3) "**x ray diffraction**", "**x ray crystallography**", "**r barm space**", "**barm space group**" (3) "**infe1 x deltatix**", "**delta doped gaas**", "**space group p**", "**crystal x ray**", "**feo6 mno6 octahedra**", "**femn2o12 infinity chains**", "**monoclinic space group**", "**c academic press**", "**space group p2**", "**mno6 octahedra chains**", "**single crystal x**", "**ir spectroscopy x**", "**spectroscopy x ray**" (2) focuses on crystal structure determination, emphasizing the use of x-ray diffraction and crystallography.

(9) Cluster 63) "**sp**" (31) "**n**" (27) "**p**", "**mexico**" (19) "**species**" (14) "**first**" (9) "**new**", "**time**", "**recorded**" (7) "**host**", "**sciadicleithrum**" (6) "**moser**" (5) "**gen**", "**panama**", "**genus**", "**helminths**", "**bates**", "**americanum**", "**helicina**" (4) "**one**", "**large**", "**h**", "**d**", "**morphology**", "**costa**", "**rica**", "**america**", "**shape**", "**specimens**", "**genera**", "**teeth**", "**nicaragua**", "**tursionis**", "**triangularis**", "**thrissina**", "**libertate**", "**psephenotarsis**", "**psepheninae**", "**groschafti**", "**micropleura**" (3) "**n sp**" (15) "**first time**" (7) "**sp n**" (6) "**recorded first**" (5) "**time mexico**" (4) "**d helicina**", "**sp p**", "**micropleura sp**", "**n gen**", "**gen n**", "**costa rica**", "**panama p**", "**n sciadicleithrum**" (3) "**recorded first time**" (5) "**first time mexico**" (4) "**n gen n**", "**sp n sciadicleithrum**", "**gen n sp**" (3) "**costa rica panama**", "**pinotepa n sp**", "**perez vigueras 1956**", "**p pachypyga burmeister**", "**groschafti sp n**", "**pseudoneodiplostomum groschafti sp**", "**n sp p**" (2) focuses on species of parasites occurring throughout Latin America.

(9) Cluster 64) "**mexico**" (25) "**mortality**" (21) "**population**", "**cancer**" (16) "**rates**" (13) "**total**", "**rate**", "**mexican**", "**health**" (12) "**birth**" (10) "**s**", "**years**" (9) "**age**", "**women**", "**childhood**" (8) "**data**", "**period**", "**increase**", "**1991**", "**states**", "**methods**", "**national**", "**city**", "**healthcare**", "**fertility**", "**twinning**" (7) "**c**", "**science**", "**family**", "**sc**", "**northeastern**", "**million**", "**000**", "**trend**", "**pedestrian**", "**asthma**", "**bronchial**" (6) "**childhood cancer**" (7) "**c science**", "**bronchial asthma**" (6) "**cancer mortality**", "**pedestrian injuries**", "**mortality rates**", "**mexico city**" (5) "**1996 million**", "**100 000**",

"**twinning rate**", "**northeastern mexico**", "**social security**" (4) "**u s**", "**fatal pedestrian**", "**mexican social**", "**family planning**", "**years age**", "**1991 1996**", "**united states**", "**rates incidence**", "**health spending**" (3) "**fatal pedestrian injuries**", "**mexican social security**" (3) focuses on macro-level studies and financial implications of diverse diseases and injuries.

## APPENDIX 7 – PARTITIONAL CLUSTERING METHOD

CLUTO (53) is a software package that implements various algorithms for clustering low- and high-dimensional datasets and for analyzing the characteristics of the various clusters. CLUTO implements three different classes of clustering algorithms that can operate either directly in the object's feature space or in the object's similarity space. The clustering algorithms provided by CLUTO are based on the partitional, agglomerative, and graph-partitioning paradigms. CLUTO's partitional and agglomerative algorithms are able to find clusters that are primarily globular, whereas its graph-partitioning and some of its agglomerative algorithms are capable of finding transitive clusters.

In this study, documents were clustered using the partitional clustering algorithms provided by CLUTO. Partitional clustering algorithms find the clusters by partitioning the entire document collection into a predetermined number of disjoint sets, each corresponding to a single cluster. This partitioning is achieved by treating the clustering process as an optimization procedure that tries to create high quality clusters according to a particular function that reflects the underlying definition of the “goodness” of the clusters. This function is referred to as the *clustering criterion function.* CLUTO implements seven such criterion functions that measure various aspects of intra-cluster similarity, inter-cluster dissimilarity, and their combinations, and have been shown to produce high-quality clusters in low- and high-dimensional datasets [Zhao and Karypis, 2003].

CLUTO uses two different methods for computing the partitioning clustering solution. The first method computes a *k*-way clustering solution via a sequence of repeated bisections, whereas the second method computes the solution directly (in a fashion similar to traditional *K*-means-based algorithms). These methods are often referred to as *repeated bisecting* and *direct k*-way clustering, respectively. CLUTO computes a direct *k*-way clustering as follows. Initially, a set of *k* objects is selected from the datasets to act as the *seeds* of the *k* clusters. Then, for each object, its similarity to these *k* seeds is computed, and it is assigned to the cluster corresponding to its most similar seed. This forms the initial *k*-way clustering. This clustering is then repeatedly refined so that it optimizes a desired clustering criterion function. This optimization is performed using a randomized incremental optimization algorithm that is greedy in nature, has low computational requirements, and produces high-quality solutions [Zhao and Karypis, 2003]. A *k*-way partitioning via repeated bisections is obtained by recursively applying the above algorithm to compute 2-way clustering (*i.e*., bisections). Initially, the objects are partitioned into two clusters, then one of these clusters is selected and is further bisected, and so on. This process continues *k* - 1 times, leading to *k* clusters. Each of these bisections is performed so that the resulting two-way clustering solution optimizes a particular criterion function.

The actual documents were represented using the widely-used vector-space model in which the various terms present in the documents were used to define a high-dimensional space and each document was considered to be a vector in that space. However, unlike the traditional vector-space representation, which relies entirely on single terms, all consecutive two- and three-word combinations were taken into account, resulting in a representation that is capable of capturing the phrases commonly occurring in the documents. In addition, Porter’s stemming algorithm was used to pre-process the various terms of each document prior to obtaining their vector-space representation. The weight of each dimension was computed using the TF-IDF model in which terms that occur many times within a document are given higher weight (TF) and terms that occur across many documents were given lower weight (IDF) [Zhao and Karypis, 2003]. The similarity between two documents was measured using the cosine of their corresponding document vectors.

## APPENDIX 8 – PARTITIONAL CLUSTERING CLUSTERS

### 34 CLUSTER RESULTS

(212) Cluster 0) (astrophi 45.8%, astron 37.8%, galaxi 2.2%, mon 1.6%, star 1.2%, astr 0.3%, suppl 0.3%, soc 0.3%, sup 0.3%, space 0.2%, astrofi 0.2%, observ 0.2%, radio 0.2%, astrophys 0.2%, nebula 0.2%, emiss 0.2%, jet 0.2%, ism 0.1%, apj 0.1%, ngc 0.1%, stellar 0.1%, veloc 0.1%, maser 0.1%, pac 0.1%, disk 0.1%, publ 0.1%, format 0.1%, region 0.1%, line 0.1%, iau 0.1%, opt 0.1%, turbul 0.1%, cluster 0.1%, telescop 0.1%, galact 0.1%, 2000 0.1%, wind 0.1%, phy 0.1%, astronomi 0.1%, sourc 0.1%) focuses on astronomy and astrophysics, emphasizing observations of different spectral emissions from stars and galaxies.

(46) Cluster 1) (polym 73.1%, polymer 1.1%, macromolecul 1.0%, blend 0.8%, graft 0.7%, pol 0.7%, poli 0.6%, vinyl 0.6%, chem 0.6%, sci 0.6%, eng 0.5%, guadalajara 0.4%, quim 0.4%, copolym 0.4%, polyethylen 0.4%, composit 0.4%, macromol 0.4%, appl 0.4%, monom 0.3%, mat 0.3%, mater 0.3%, methacryl 0.3%, makromol 0.2%, 25100 0.2%, acryl 0.2%, particl 0.2%, 44430 0.2%, properti 0.2%, emuls 0.2%, foam 0.1%, valladolid 0.1%, saltillo 0.1%, rheol 0.1%, coahuila 0.1%, ldpe 0.1%, hydrogel 0.1%, film 0.1%, bull 0.1%, plast 0.1%, ingn 0.1%) focuses on macromolecular polymers, emphasizing polymerization of vinyl monomers and the synthesis of graft co-polymers.

(57) Cluster 2) (catal 60.3%, catalyst 4.5%, chem 2.3%, zeolit 1.5%, al2o3 0.9%, alumina 0.8%, catalysi 0.8%, catalyt 0.5%, bokhimi 0.5%, petr 0.5%, appl 0.5%, titania 0.5%, mater 0.4%, 07730 0.4%, gel 0.4%, quim 0.3%, hydrotr 0.3%, mexicano 0.3%, vu 0.3%, surfac 0.3%, boehmit 0.3%, solid 0.3%, gen 0.3%, oil 0.3%, sol 0.3%, reaction 0.3%, eng 0.2%, clai 0.2%, phy 0.2%, tio2 0.2%, micropor 0.2%, phase 0.2%, crystallin 0.2%, genoa 0.2%, ej 0.2%, ind 0.2%, mat 0.2%, chim 0.2%, acid 0.2%, crystallit 0.2%) - focuses on chemical catalysis, emphasizing zeolite-alumina catalyts.

(52) Cluster 3) geophi 38.7%, seismol 7.8%, wave 6.0%, planet 2.4%, space 2.1%, magnet 1.6%, earthquak 1.6%, geofi 1.6%, earth 1.6%, geophys 1.4%, re 1.4%, field 0.7%, acoust 0.5%, seismic 0.5%, plasma 0.4%, russel 0.4%, motion 0.4%, slip 0.4%, shock 0.4%, frequenc 0.4%, soc 0.4%, atmo 0.3%, paleointens 0.3%, fluid 0.3%, cices 0.3%, lett 0.3%, cyclotron 0.3%, sol 0.3%, ion 0.3%, fault 0.3%, goguitchaichvili 0.3%, ensenada 0.2%, phy 0.2%, 04510 0.2%, mode 0.2%, autonom 0.2%, propag 0.2%, solar 0.2%, ground 0.2%, huddleston 0.2%) - focuses on analysis of waves from seismic instrumentation, emphasizing passive detection signals from catastrophic geophysical events such as earthquakes, but including active detection signals from seismic signals for exploration and archaeology.

(68) Cluster 4) (wheat 20.2%, crop 13.0%, cimmyt 6.7%, genet 6.5%, 06600 4.3%, maiz 3.8%, plant 3.8%, breed 2.2%, 641 1.5%, improv 1.4%, agr 1.2%, euphytica 0.9%, agron 0.8%, mergoum 0.8%, yield 0.8%, theor 0.7%, rajaram 0.7%, genotyp

0.6%, resist  0.6%, int  0.6%, grain  0.5%, cereal  0.5%, marker  0.4%, germplasm  0.4%, gene  0.4%, cultivar  0.4%, genom  0.4%, sci  0.4%, line  0.4%, lisboa  0.3%, registr 0.3%, popul  0.3%, trait  0.3%, bread  0.3%, tritical  0.3%, chromosom  0.3%, durum 0.3%, triticum  0.3%, 445qx  0.2%, kazi  0.2%) – focuses on wheat and allied crops such as maize, emphasizing genetic modification for improved yield and benefits of storage in gene banks such as CIMMYT.

(76) Cluster 5) (bacteriol 16.7%, microbiol 12.9%, gene  9.9%, mol  8.4%, biol  2.2%, coli  2.0%, protein  2.0%, genet  1.9%, escherichia  1.4%, nucleic  1.2%, acid  0.8%, transcript  0.8%, cuernavaca  0.8%, biochem  0.8%, express  0.8%, sequenc  0.7%, morelo  0.7%, acad  0.7%, regul  0.6%, microb  0.6%, mutant  0.6%, operon  0.6%, natl 0.6%, biotecnol  0.6%, fem  0.5%, biotechnol  0.5%, usa  0.5%, cell  0.5%, clone  0.4%, plant  0.3%, embo  0.3%, strain  0.3%, toxin  0.3%, chem  0.3%, encod  0.3%, activ 0.3%, 510  0.3%, promot  0.3%, v179  0.3%, infect  0.3%) - focuses on bacteriology and microbiology, emphasizing genetic and protein studies.

(163) Cluster 6) (opt 48.2%, optic 13.4%, puebla  2.1%, laser  1.5%, fiber  1.2%, elect 1.2%, lett  1.1%, 72000  1.1%, phy  1.1%, commun  0.9%, wave  0.7%, nonlinear  0.7%, grate  0.6%, appl  0.6%, ieee  0.6%, soliton  0.6%, beam  0.5%, photo  0.5%, astrofi 0.5%, fi  0.4%, leon  0.4%, soc  0.4%, 37150  0.4%, ensenada  0.4%, 216  0.4%, imag 0.4%, gto  0.4%, photorefract  0.3%, quantum  0.3%, puls  0.3%, in  0.3%, ctr  0.3%, sensor  0.2%, benemerita  0.2%, diffract  0.2%, photon  0.2%, polar  0.2%, propag  0.2%, birefring  0.2%, crystal  0.2%) – focuses on laser-fiber optic interactions.

(93) Cluster 7) (nucl 13.8%, phy 12.4%, usa  4.9%, instrum  3.7%, radiat  2.4%, meth 1.6%, univ  1.0%, lett  0.9%, in2p3  0.9%, abbott  0.9%, germani  0.8%, detector  0.7%, mea  0.7%, radon  0.7%, russia  0.6%, track  0.6%, fi  0.6%, franc  0.5%, itali  0.5%, rev 0.4%, cnr  0.4%, pari  0.4%, dosim  0.4%, calif  0.4%, wuppert  0.4%, adloff  0.4%, janeiro  0.4%, davi  0.4%, berkelei  0.4%, ion  0.3%, measur  0.3%, gamma  0.3%, moscow  0.3%, republ  0.3%, decai  0.3%, fermilab  0.3%, brazil  0.3%, energi  0.3%, aitala  0.3%, czech  0.3%) – focuses on nuclear physics, emphasizing radiation detection instrumentation.

(86) Cluster 8) (geol 31.1%, geologi  4.9%, geophi  3.1%, earth  2.8%, geolog  2.6%, volcanol  2.4%, sediment  1.7%, miner  1.4%, geotherm  1.4%, volcan  1.3%, bull  1.2%, basin  1.1%, geochim  0.9%, geoth  0.9%, tecton  0.9%, cosmochim  0.9%, cretac  0.8%, soc  0.7%, aapg  0.7%, geochem  0.7%, geofi  0.6%, rock  0.6%, california  0.6%, isotop 0.5%, planet  0.5%, re  0.5%, paleontol  0.5%, volcano  0.4%, verma  0.4%, fault  0.4%, magma  0.4%, baja  0.3%, petrol  0.3%, sedimentari  0.3%, contrib  0.3%, erupt  0.3%, mar  0.3%, water  0.3%, basalt  0.2%, zone  0.2%) – focuses on geology and geophysics, emphasizing volcanology and study of water basin sediments.

(68) Cluster 9) (eng 15.9%, chem  8.5%, petr  3.7%, 07730  3.2%, ind  3.0%, aich  2.8%, asphalten  2.6%, mexicano  2.5%, fluid  2.1%, oil  1.5%, equilibr  1.4%, ej  1.4%, phase 1.4%, 152  1.2%, lazaro  1.1%, control  1.1%, liquid  1.1%, cardena  1.0%, fuel  0.9%, cent  0.9%, colloid  0.9%, mixtur  0.8%, equilibria  0.7%, porou  0.7%, reactor  0.6%,

langmuir 0.6%, distil 0.5%, metropolitana 0.5%, petrol 0.5%, vapor 0.5%, simulac 0.5%, engn 0.4%, data 0.4%, interf 0.4%, technol 0.4%, iztapalapa 0.4%, pressur 0.4%, separ 0.4%, process 0.3%, mayagoitia 0.3%) – focuses on petroleum engineering, emphasizing petroleum chemistry.

(66) Cluster 10) (biotechnol 27.4%, bioeng 8.5%, microb 5.2%, microbiol 4.8%, biot 2.7%, ferment 2.6%, enzym 1.9%, biotecnol 1.7%, oil 1.4%, cultur 1.1%, appl 1.1%, environ 1.0%, product 1.0%, chem 0.6%, algin 0.6%, tech 0.5%, biotechnolog 0.5%, yeast 0.5%, technol 0.5%, phenol 0.4%, biochem 0.4%, eng 0.4%, bioingn 0.4%, strain 0.4%, laccas 0.4%, medium 0.4%, bioreactor 0.3%, astaxanthin 0.3%, immobil 0.3%, glucos 0.3%, progr 0.3%, acid 0.3%, membran 0.3%, bagass 0.3%, biomass 0.3%, chitosan 0.3%, ga 0.3%, bioprocess 0.3%, concentr 0.3%, morelo 0.2%) – focuses on biotechnology and bioengineering, emphasizing enzymatic reactions and microbial processes in fermentation.

(94) Cluster 11) (fish 23.0%, aquacultur 13.7%, mar 4.4%, shrimp 4.1%, biol 2.6%, fisheri 1.7%, aquacult 1.6%, shellfish 1.4%, paz 1.3%, ecol 1.1%, aquat 1.0%, baja 0.9%, world 0.9%, marina 0.9%, diet 0.9%, california 0.6%, growth 0.5%, 23000 0.5%, noroest 0.4%, marin 0.4%, scallop 0.4%, feed 0.4%, calif 0.4%, vannamei 0.4%, pesquera 0.3%, salin 0.3%, oyster 0.3%, digest 0.3%, fao 0.3%, cienc 0.3%, thesi 0.3%, noaa 0.3%, ciencia 0.3%, merida 0.3%, larva 0.3%, gulf 0.3%, protein 0.3%, ctr 0.3%, speci 0.3%, biochem 0.2%) – focuses on fish and marine aquaculture, emphasizing shrimp and other shellfish.

(158) Cluster 12) (film 25.4%, thin 7.2%, electrochem 4.2%, solar 3.9%, solid 3.2%, sol 3.1%, phy 2.4%, deposit 2.3%, mater 1.5%, appl 1.3%, mat 1.2%, energ 1.1%, energi 0.9%, coat 0.9%, 62580 0.9%, temixco 0.8%, nair 0.8%, ipn 0.7%, temperatur 0.7%, hydrogen 0.7%, anneal 0.6%, oxid 0.6%, substrat 0.6%, cell 0.5%, chemic 0.5%, cryst 0.5%, surf 0.4%, growth 0.4%, fi 0.4%, morelo 0.4%, chem 0.4%, degreesc 0.4%, vac 0.4%, cdte 0.4%, queretaro 0.3%, materi 0.3%, electron 0.3%, lett 0.3%, electrod 0.3%, energia 0.3%) – focuses on thin films, emphasizing solar applications such as protective coating and electrical conversion.

(123) Cluster 13) (immunol 16.2%, parasitol 8.6%, vet 8.0%, infect 7.8%, immun 5.9%, cell 4.5%, med 2.5%, parasit 2.5%, antigen 1.5%, microbiol 1.3%, virol 1.2%, antibodi 1.1%, mice 0.8%, viru 0.7%, vaccin 0.7%, biol 0.6%, exp 0.5%, anim 0.5%, lymphocyt 0.4%, biom 0.4%, protein 0.4%, pig 0.4%, respons 0.4%, mol 0.4%, immunolog 0.4%, cysticercosi 0.4%, clin 0.4%, di 0.3%, taenia 0.3%, human 0.3%, trop 0.3%, parasitolog 0.3%, blood 0.3%, solium 0.3%, pathol 0.3%, 04510 0.3%, express 0.2%, biochem 0.2%, dairi 0.2%, re 0.2%) – focuses on immunology at the cellular level, and parasitology to study animal diseases.

(422) Cluster 14) (phy 55.1%, rev 5.3%, lett 4.1%, fi 2.3%, nucl 2.1%, math 1.3%, quantum 1.1%, chem 0.7%, physic 0.6%, mod 0.5%, physica 0.5%, theori 0.4%, field 0.3%, classic 0.3%, ciencia 0.3%, model 0.3%, atom 0.3%, autonoma 0.2%, opt 0.2%, energi 0.2%, mol 0.2%, 2000 0.2%, equat 0.2%, gen 0.2%, 01000 0.2%, theor

0.2%, nojiri  0.2%, postal  0.2%, nacl  0.2%, univ  0.1%, quant  0.1%, dynam  0.1%, apartado  0.1%, grav  0.1%, particl  0.1%, gravit  0.1%, system  0.1%, electron  0.1%, inst  0.1%, neutrino  0.1%) – focuses on physics articles in many areas of physics, especially articles published in physical review letters.

(159) Cluster 15) (neurosci 18.5%, brain 17.0%, rat  3.7%, neuron  2.9%, behav  2.8%, physiol  2.5%, neurochem  2.5%, neurol  1.9%, re  1.2%, cell  1.1%, neurobiol  1.1%, fisiol  1.1%, sleep  0.9%, pharmacol  0.9%, anim  0.8%, receptor  0.7%, exp  0.7%, psychol  0.6%, neurociencia  0.5%, comp  0.5%, biol  0.5%, neuroreport  0.5%, activ 0.4%, biochem  0.4%, taurin  0.4%, med  0.4%, endocrinolog  0.3%, learn  0.3%, neuroendocrinolog  0.3%, gaba  0.3%, neurophysiol  0.3%, natl  0.3%, induc  0.3%, hormon  0.3%, behavior  0.3%, dopamin  0.2%, neuro  0.2%, releas  0.2%, acad  0.2%, trh  0.2%) – focuses on neuroscience, concentrating on neuron activity in the brain, and on studies in the rat.

(129) Cluster 16) (food 45.0%, phytochemistri  3.9%, agr  2.8%, chem  1.9%, prod  1.5%, cereal  1.2%, quim  1.1%, anim  1.0%, sci  0.8%, extract  0.7%, technol  0.6%, alimento 0.6%, tortilla  0.6%, acid  0.5%, starch  0.5%, corn  0.5%, fruit  0.5%, dry  0.5%, plant 0.5%, dairi  0.4%, nutr  0.4%, degreesc  0.4%, protein  0.3%, nat  0.3%, oil  0.3%, queretaro  0.3%, autonoma  0.3%, activ  0.3%, storag  0.3%, alkaloid  0.3%, biochem 0.3%, meat  0.2%, product  0.2%, planta  0.2%, fac  0.2%, microbiol  0.2%, tecnol  0.2%, water  0.2%, cesped  0.2%, milk  0.2%) – focuses on food agriculture, emphasizing cereal chemistry.

(133) Cluster 17) (diabet 12.2%, nutr  9.6%, clin  6.6%, med  6.2%, metab  3.4%, women 2.7%, insulin  2.7%, obstet  2.2%, endocr  2.2%, gynecol  1.9%, obes  1.8%, endocrinol 1.3%, reprod  1.1%, zubiran  0.7%, care  0.7%, epidemiol  0.7%, salvador  0.6%, serum 0.6%, subject  0.6%, contracept  0.5%, 14000  0.5%, patient  0.5%, diet  0.4%, metabol 0.4%, ag  0.4%, glucos  0.4%, social  0.4%, invest  0.4%, level  0.4%, cholesterol  0.4%, hormon  0.4%, bmi  0.3%, quiroga  0.3%, vasco  0.3%, hum  0.3%, lipid  0.3%, re  0.3%, mutat  0.3%, risk  0.3%, blood  0.3%) – focuses on nutritional approaches to precent diabetes, and clinical approaches to treat diabetes.

(148) Cluster 18) (ieee 26.6%, control  7.4%, contr  3.3%, robot  3.3%, automat  3.2%, syst  2.5%, power  2.3%, circuit  2.0%, elect  1.9%, system  1.8%, network  1.2%, comput  1.1%, neural  1.0%, algorithm  0.9%, automatica  0.8%, int  0.7%, nonlinear 0.7%, ingn  0.6%, robust  0.6%, ipn  0.6%, engn  0.6%, design  0.5%, comp  0.5%, chao 0.4%, linear  0.4%, tran  0.4%, cinvestav  0.4%, dynam  0.4%, signal  0.4%, acm  0.3%, autom  0.3%, stabil  0.3%, feedback  0.3%, adapt  0.3%, math  0.3%, optim  0.3%, model 0.3%, approach  0.3%, base  0.3%, simul  0.2%) – focuses on automatic control of robots.

(131) Cluster 19) (entomol 37.7%, speci  5.6%, zool  3.7%, parasitol  2.8%, biol  2.0%, brailovski  1.2%, parasit  1.1%, ecol  1.0%, insect  1.0%, wash  0.7%, entomolog  0.7%, host  0.6%, moravec  0.6%, bee  0.6%, folia  0.6%, genu  0.6%, soc  0.5%, yucatan 0.5%, 04510  0.5%, mosquito  0.4%, femal  0.4%, postal  0.4%, 70153  0.4%, fish  0.4%, mu  0.3%, 153  0.3%, new  0.3%, veracruz  0.3%, coreida  0.3%, male  0.3%,

entomologist 0.3%, colima 0.3%, fla 0.3%, 91000 0.3%, xalapa 0.3%, helminth 0.3%, chiapa 0.3%, econ 0.3%, oviposit 0.2%, republ 0.2%) – focuses on zoological entomological species.

(119) Cluster 20) (pharmacol 25.0%, physiol 5.1%, rat 2.7%, pharm 2.6%, alpha 2.0%, receptor 1.9%, biochem 1.6%, drug 1.5%, ther 1.2%, eur 1.2%, adrenoceptor 1.1%, ca2 1.0%, toxicol 0.9%, med 0.9%, liver 0.9%, cell 0.9%, brit 0.8%, channel 0.8%, biol 0.7%, induc 0.7%, effect 0.6%, clin 0.5%, muscl 0.5%, arch 0.5%, exp 0.5%, hypertens 0.4%, activ 0.4%, ipn 0.4%, re 0.4%, inhibit 0.4%, chem 0.4%, london 0.4%, pharmacolog 0.3%, biophi 0.3%, brain 0.3%, respons 0.3%, dose 0.3%, mol 0.3%, hepatolog 0.3%, carotid 0.3%) – focuses on pharmacological studies of drugs in animals.

(155) Cluster 21) (plant 32.3%, soil 14.7%, bot 2.3%, physiol 2.2%, cell 1.4%, biol 1.4%, seed 1.3%, agr 1.0%, microbiol 0.9%, environ 0.7%, germin 0.7%, crop 0.7%, ecol 0.7%, biochem 0.6%, mol 0.6%, planta 0.5%, microb 0.5%, root 0.5%, genet 0.5%, irapuato 0.5%, phyton 0.4%, hortic 0.4%, sci 0.4%, protein 0.4%, gene 0.4%, bacteriol 0.3%, syst 0.3%, plantarum 0.3%, agron 0.3%, fruit 0.3%, nitrogen 0.2%, growth 0.2%, ecolog 0.2%, phytopatholog 0.2%, eros 0.2%, hortscienc 0.2%, cultur 0.2%, irrig 0.2%, usa 0.2%, yucatan 0.2%) – focuses on botanical studies of plants and their soils.

(277) Cluster 22) (chem 35.5%, quim 5.8%, tetrahedron 3.8%, inorg 3.6%, org 1.6%, soc 1.4%, phy 1.4%, acta 1.1%, complex 1.0%, organomet 1.0%, nmr 1.0%, crystallogr 0.9%, 04510 0.8%, crystal 0.7%, chim 0.7%, structur 0.7%, heterocycl 0.6%, compound 0.5%, synthesi 0.4%, exterior 0.4%, circuito 0.4%, reaction 0.4%, reson 0.4%, coordin 0.4%, polyhedron 0.4%, mol 0.4%, magn 0.3%, ligand 0.3%, bond 0.3%, organometal 0.3%, basiuk 0.3%, sheldrick 0.3%, theochem 0.3%, autonoma 0.3%, dalton 0.3%, anal 0.3%, asymmetr 0.3%, chemistri 0.3%, citi 0.3%, rai 0.3%) – focuses on structural chemistry

(144) Cluster 23) surg 21.7%, patient 8.9%, oncol 5.0%, dermatol 4.3%, cancer 2.8%, med 2.4%, clin 2.1%, ophthalmol 1.9%, surgeri 1.6%, plast 1.2%, pathol 1.2%, hosp 1.1%, tumor 0.8%, gastroenterol 0.7%, lymphoma 0.6%, hematol 0.6%, clinic 0.6%, diseas 0.5%, case 0.5%, di 0.5%, oral 0.5%, blood 0.5%, pediatr 0.5%, chemotherapi 0.5%, leukemia 0.5%, thorac 0.5%, cell 0.5%, pediat 0.4%, brit 0.4%, reconstr 0.4%, gastroenterolog 0.4%, month 0.4%, arch 0.4%, treatment 0.3%, carcinoma 0.3%, surgic 0.3%, therapi 0.3%, 14000 0.3%, ophthalmolog 0.3%, ey 0.3% - focuses on clinical medicine, especially cancer surgery.

(171) Cluster 24) mar 15.0%, california 5.1%, baja 3.3%, chetum 2.4%, biol 2.3%, marin 2.1%, limnol 2.0%, ecol 1.8%, mazatlan 1.8%, gulf 1.6%, fish 1.5%, ocean 1.3%, reef 1.3%, environ 1.3%, speci 1.3%, ensenada 1.2%, pollut 0.9%, hydrobiologia 0.9%, oceanogr 0.9%, 77000 0.9%, coral 0.8%, sea 0.8%, cienc 0.6%, water 0.6%, coast 0.6%, abund 0.6%, sur 0.6%, roo 0.6%, zool 0.6%, marina 0.6%, trop 0.6%, crustaceana 0.5%, frontera 0.5%, bull 0.5%, oceanol 0.5%, quintana 0.4%,

ser  0.4%, paz  0.4%, pacif  0.4%, caribbean  0.4% - focuses on marine ecology, especially in Baja California region.

(145) Cluster 25) patient  7.3%, genet  6.5%, med  6.3%, cardiol  4.9%, rheumatol  4.6%, arthriti  4.3%, rheum  3.8%, clin  2.6%, hum  2.4%, immunol  1.7%, hosp  1.5%, kidnei 1.5%, 14080  1.0%, neurol  1.0%, ignacio  1.0%, lupu  1.0%, transplant  0.9%, diseas 0.8%, di  0.8%, autoimmun  0.7%, heart  0.7%, nephrol  0.6%, hla  0.6%, chavez  0.6%, dialysi  0.6%, allel  0.5%, hypertens  0.5%, syndrom  0.5%, echocardiographi  0.5%, periton  0.4%, antigen  0.4%, sle  0.4%, chromosom  0.4%, neurolog  0.4%, badiano 0.4%, echocardiog  0.4%, engl  0.4%, imss  0.3%, nacl  0.3%, citi  0.3% - focuses on clinical medicine, especially genetic and autoimmune diseases.

(217) Cluster 26) math 41.2%, matemat  6.3%, algebra  5.9%, oper  1.1%, theori  1.1%, space  0.9%, equat  0.8%, topolog  0.7%, appl  0.7%, stat  0.6%, note  0.6%, mech  0.6%, lect  0.5%, anal  0.5%, integr  0.4%, commun  0.4%, comput  0.4%, function  0.4%, phy 0.3%, discret  0.3%, siam  0.3%, mat  0.3%, soc  0.3%, dokl  0.3%, geom  0.3%, optim 0.3%, mathemat  0.3%, ann  0.3%, finit  0.3%, autonoma  0.3%, linear  0.3%, set  0.3%, method  0.3%, numer  0.2%, theor  0.2%, nonlinear  0.2%, dept  0.2%, map  0.2%, morelia  0.2%, solut  0.2% - focuses on theoretical and applied mathematics.

(141) Cluster 27) mater 19.4%, ceram  9.0%, mat  3.1%, alloi  2.7%, metal  2.7%, saltillo 2.0%, materi  1.9%, glass  1.7%, magn  1.5%, powder  1.5%, phase  1.4%, solid  1.1%, corros  1.1%, ferroelectr  0.8%, met  0.8%, temperatur  0.8%, phy  0.6%, cryst  0.6%, coahuila  0.6%, degreesc  0.5%, microstructur  0.5%, chem  0.5%, gel  0.4%, hydroxyapatit  0.4%, magnet  0.4%, sci  0.4%, invest  0.4%, queretaro  0.4%, particl 0.4%, electrochem  0.3%, crystal  0.3%, ipn  0.3%, coat  0.3%, fi  0.3%, diffract  0.3%, solut  0.3%, mechan  0.3%, properti  0.3%, process  0.3%, oxid  0.3% - focuses on materials, especially ceramics.

(155) Cluster 28) cell 16.2%, biol  6.4%, biochem  4.7%, mutat  2.8%, mol  2.6%, reprod 1.7%, protein  1.6%, biochemistri  1.5%, biophi  1.5%, chem  1.4%, activ  1.2%, cancer 1.1%, genet  1.0%, toxicol  1.0%, med  0.9%, re  0.9%, sperm  0.8%, express  0.7%, mutagen  0.7%, dev  0.7%, natl  0.7%, dna  0.7%, endocrinolog  0.6%, physiol  0.5%, receptor  0.5%, beta  0.4%, biochim  0.4%, acad  0.4%, usa  0.4%, celular  0.4%, apoptosi 0.4%, gene  0.4%, bind  0.4%, endocrinol  0.4%, membran  0.3%, natur  0.3%, induc 0.3%, bioquim  0.3%, acid  0.3%, enzym  0.3% - focuses on cell biology and biochemistry.

(100) Cluster 29) flow  9.5%, heat  8.1%, eng  5.2%, fluid  4.9%, steel  2.4%, mech 2.0%, tran  1.7%, isij  1.5%, asc  1.1%, ingn  1.1%, engn  1.1%, turbul  1.0%, int  0.9%, met  0.9%, hidraul  0.9%, model  0.9%, comput  0.8%, transfer  0.8%, metal  0.8%, numer  0.8%, hydraul  0.7%, earthquak  0.6%, rheol  0.6%, tundish  0.5%, convect  0.5%, asm  0.5%, channel  0.5%, thermal  0.5%, mass  0.5%, process  0.5%, steelmak  0.4%, morelo  0.4%, esiqi  0.4%, design  0.4%, colloid  0.4%, equat  0.4%, temixco  0.4%, tecnol  0.4%, flotat  0.4%, proc  0.4% - focuses on heat and fluid flow.

(170) Cluster 30) phy 23.1%, appl 4.5%, puebla 2.8%, solid 2.2%, fi 2.2%, lett 2.1%, gaa 1.6%, electron 1.6%, film 1.5%, semiconductor 1.3%, rev 1.2%, layer 1.0%, optic 1.0%, solidi 0.9%, fluid 0.7%, statu 0.6%, surfac 0.6%, mater 0.6%, mat 0.6%, crystal 0.6%, cryst 0.5%, epitaxi 0.5%, growth 0.5%, physic 0.5%, 72570 0.5%, opt 0.5%, thermal 0.4%, laser 0.4%, quantum 0.4%, sov 0.4%, ipn 0.4%, thin 0.4%, vac 0.4%, dope 0.3%, energi 0.3%, magnet 0.3%, autonoma 0.3%, field 0.3%, mod 0.3%, photoluminesc 0.3% - focuses on solid state physics, especially semiconductors with films.

(106) Cluster 31) environ 13.4%, water 12.8%, atmo 5.1%, air 2.7%, manag 2.5%, wast 2.5%, technol 1.9%, pollut 1.6%, sludg 1.1%, atmosfera 1.0%, anaerob 0.7%, agua 0.7%, aerosol 0.6%, env 0.6%, tox 0.6%, contam 0.6%, emiss 0.6%, soil 0.5%, mex 0.5%, hidraul 0.5%, epa 0.4%, meteorol 0.4%, concentr 0.4%, ambient 0.4%, ozon 0.4%, toxic 0.4%, ing 0.4%, wastewat 0.4%, re 0.4%, sci 0.3%, model 0.3%, qualiti 0.3%, wat 0.3%, hydrolog 0.3%, drink 0.2%, health 0.2%, sediment 0.2%, int 0.2%, remov 0.2%, environment 0.2% - focuses on ecological management of the environment, especially the impact of pollution on air and water.

(163) Cluster 32) ecol 11.5%, bot 8.9%, forest 6.9%, ecolog 4.2%, speci 3.8%, biol 2.5%, conserv 2.0%, mammal 1.1%, plant 1.0%, chiapa 1.0%, bat 0.9%, biotropica 0.9%, nat 0.9%, manag 0.8%, zool 0.8%, habitat 0.8%, tree 0.8%, veracruz 0.7%, tropic 0.7%, bird 0.7%, popul 0.6%, evol 0.5%, divers 0.5%, evolut 0.5%, xalapa 0.5%, land 0.5%, syst 0.5%, oecologia 0.4%, 91000 0.4%, wildlif 0.4%, flora 0.4%, trop 0.4%, season 0.4%, liana 0.4%, fruit 0.3%, thesi 0.3%, biodivers 0.3%, mycologia 0.3%, mycotaxon 0.3%, autonoma 0.3% - focuses on ecological management and conservation, especially forests, plants, and animals.

(179) Cluster 33) infect 12.1%, med 9.3%, di 5.3%, clin 3.4%, health 3.3%, hlth 2.0%, microbiol 1.9%, salud 1.4%, hosp 1.4%, cancer 1.3%, care 1.3%, children 1.0%, epidemiol 0.8%, patient 0.7%, pediatr 0.7%, lead 0.7%, antimicrob 0.6%, tuberculosi 0.5%, arch 0.5%, publ 0.5%, risk 0.5%, alcohol 0.5%, resist 0.5%, lung 0.5%, resp 0.5%, usa 0.5%, isol 0.4%, women 0.4%, drug 0.4%, hpv 0.4%, imss 0.4%, clinic 0.4%, medic 0.4%, lancet 0.4%, virol 0.3%, pediat 0.3%, engl 0.3%, respir 0.3%, blood 0.3%, environ 0.3% - focuses on infectious disease clinical medicine.

## 64 CLUSTER RESULTS

(37) Cluster 0) (fruit 26.6%, seed 18.2%, flower 9.5%, germin 5.0%, plant 1.2%, alkaloid 0.8%, pollin 0.5%, harvest 0.5%, degreesc 0.5%, product 0.5%) – focuses on germination of fruit and flower seeds.

(37) Cluster 1) (receptor 17.9%, alpha 6.0%, adrenoceptor 5.2%, alpha.adrenoceptor 4.7%, antagonist 2.7%, dopamin 2.3%, mug 1.5%, carotid 1.2%, respons 1.1%, induc 1.1%) – focuses on biological effects of receptors and adrenoreceptors.

(45) Cluster 2) (ca2 27.5%, channel 6.4%, sperm 6.3%, current 2.6%, ca2.channel 1.8%, mum 1.6%, intracellular 1.4%, inhibit 1.3%, pkc 1.0%, type.ca2 0.9%) – focuses on currents in calcium channels, including sperm modulation.

(34) Cluster 3) (soil 37.1%, weed 2.0%, eros 1.7%, crop 1.5%, bean 1.3%, tillag 1.3%, maiz 0.9%, co2 0.9%, runoff 0.7%, microplot 0.7%)  - focuses on the effects of soil treatments on weeds and erosion, and eventually on crop growth and yield.

(112) Cluster 4 (film 47.7%, thin.film 3.8%, deposit 3.7%, thin 3.5%, substrat 2.1%, anneal 1.6%, temperatur 1.2%, degreesc 0.8%, optic 0.7%, film.deposit 0.7%) – focuses on thin film deposition.

(49) Cluster 5 (insulin 11.5%, women 5.8%, diabet 4.9%, obes 4.0%, bmi 3.4%, cholesterol 2.7%, mmol 2.3%, men 2.0%, fat 1.9%, group 1.3%) – focuses on relation of insulin to diabetes and obesity in women.

(39) Cluster 6) (wave 43.9%, frequenc 1.6%, nonlinear 1.2%, mode 1.1%, surfac 1.1%, stack 1.0%, dispers 0.9%, propag 0.9%, cyclotron 0.8%, ion 0.8%) – focuses on wave frequencies and nonlinear modes.

(78) Cluster 7) (galaxi 40.4%, star 2.6%, seyfert 1.5%, stellar 1.3%, dwarf 1.1%, ngc 1.1%, starburst 1.1%, spiral 1.0%, ga 1.0%, mass 1.0%) – focuses on studies of stars in galaxies.

(41) Cluster 8) (women 23.0%, hpv 7.8%, cervic 2.6%, men 1.9%, risk 1.9%, contracept 1.7%, infect 1.6%, estradiol 1.3%, sexual 1.1%, cervic.cancer 1.1%) –focuses on hpv in women and its role in cervical cancer.

(45) Cluster 9) (catalyst 21.4%, al2o3 3.8%, zeolit 3.5%, catalyt 2.7%, reaction 1.9%, xylen 1.6%, oil 1.2%, support 1.2%, la2o3 1.1%, activ 1.0%) – focuses on alumina catalysts.

(46) Cluster 10) (genotyp 10.2%, wheat 8.3%, resist 3.7%, genet 3.2%, cultivar 3.0%, line 2.7%, marker 2.5%, popul 2.4%, yield 2.2%, grain 2.2%) – focuses on resistance to infection of different wheat genotypes.

(46) Cluster 11) (grate 19.8%, beam 9.1%, photorefract 4.3%, optic 3.7%, propag 2.8%, diffract 2.1%, wave 1.8%, soliton 1.8%, nonlinear 1.7%, crystal 1.0%) – focuses on diffraction of beams on gratings through photorefractive materials.

(57) Cluster 12) (magnet 30.8%, magnet.field 9.7%, field 8.7%, plasma 1.0%, flux 0.7%, paleointens 0.6%, polar 0.5%, electron 0.5%, densiti 0.5%, approxim 0.5%) – focuses on plasmas in magnetic fields.

(86) Cluster 13) (speci 18.8%, new.speci 18.3%, new 12.0%, genu 4.3%, mexico 2.9%, illustr 2.3%, nov 1.5%, kei 1.1%, record 0.9%, collect 0.8%) – focuses on collection and identification of new species and new genus in different regions of Mexico.

(45) Cluster 14) (layer 5.6%, film 5.3%, gaa 4.9%, coat 4.5%, substrat 2.1%, epitaxi 2.1%, hydrogen 1.9%, deposit 1.8%, carbon 1.5%, oxid 1.3%) – focuses on film layers and coatings on GaA substrates.

(65) Cluster 15) (cross 10.6%, cross.section 10.5%, section 7.8%, energi 3.9%, measur 2.0%, scatter 1.9%, proton 1.7%, gev 1.3%, jet 1.1%, state 0.9%) – focuses on measurement of energy scattering cross sections.

(45) Cluster 16) (thermal 10.2%, heat 7.0%, conduct 3.6%, temperatur 3.6%, thermal.conduct 2.8%, diffus 2.3%, thermal.diffus 2.0%, solar 1.5%, photoacoust 1.4%, lifetim 1.4%) – focuses on thermal conduction and diffusion of heat.

(61) Cluster 17) (immun 9.3%, mice 8.5%, infect 8.2%, cell 3.5%, viru 2.8%, tnf 2.5%, vaccin 2.0%, parasit 1.8%, respons 1.7%, immun.respons 1.3%)  - focuses on immunity to infection experiments in mice.

(41) Cluster 18) (scatter 10.4%, surfac 6.1%, reson 4.5%, rough 2.8%, optic 2.7%, exciton 2.0%, arrai 1.6%, plane 1.5%, quantum 1.1%, polariton 1.0%) – focuses on resonant scattering from surfaces.

(58) Cluster 19) (decai 10.6%, neutrino 10.3%, gamma 3.1%, bar 1.8%, detector 1.7%, gev 1.7%, fermilab 1.6%, search 1.6%, mass 1.4%, lambda 1.4%) – focuses on decay and detection of high energy neutrinos.

(70) Cluster 20) (brane 10.6%, scalar 3.2%, phi 2.7%, cosmolog 2.2%, inflat 2.1%, graviti 2.1%, univers 2.0%, field 2.0%, metric 1.7%, einstein 1.5%) – focuses on brane inflation and cosmology with scalar fields.

(75) Cluster 21) (algebra 31.1%, oper 4.5%, polynomi 4.0%, finit 3.7%, let 1.6%, symbol 1.0%, circl 0.9%, ideal 0.8%, irreduc 0.8%, extens 0.7%) – focuses on algebras, operators, and polynomials.

(67) Cluster 22) (complex 10.4%, crystal 6.0%, angstrom 4.0%, structur 3.4%, ligand 3.3%, coordin 3.3%, rai 2.9%, macrocycl 1.7%, atom 1.6%, eta 1.5%) – focuses on crystallography of complexes and resultant angstrom-level structures.

(43) Cluster 23) (tortilla 10.5%, corn 7.4%, starch 3.8%, moistur 2.8%, nixtam 2.7%, dry 2.1%, flour 1.7%, content 1.5%, product 1.5%, medium 1.3%) - focuses on processing of grains and relation to final product quality.

(56) Cluster 24) (emiss 4.7%, absorpt 4.5%, crystal 3.6%, ion 3.3%, dope 3.0%, spectra 2.9%, band 2.8%, sampl 2.5%, excit 2.2%, luminesc 1.8%) – focuses on emission and absorption in crystals, especially ion-doped.

(100) Cluster 25) (gene 26.1%, transcript 3.3%, sequenc 2.8%, express 2.7%, mutat 2.5%, mutant 2.4%, strain 2.0%, protein 2.0%, regul 1.5%, encod 1.4%) – focuses on gene transcripts, emphasizing sequencing and expression.

(77) Cluster 26) (speci 24.6%, mexico 3.2%, collect 2.5%, fish 2.5%, record 2.3%, abund 1.9%, helminth 1.8%, yucatan 1.3%, specimen 1.2%, pacif 1.0%) - focuses on collection and recording of species of fish from different regions of Mexico.

(61) Cluster 27) (flow 11.5%, fluid 11.0%, shear 2.2%, vortic 2.1%, numer 1.6%, turbul 1.6%, porou 1.2%, tundish 1.2%, model 1.2%, water 1.1%) – focuses on fluid flow with shear and vorticity.

(62) Cluster 28) (diet 11.4%, food 6.3%, growth 4.2%, fed 2.6%, feed 2.3%, popul.growth 1.8%, larva 1.6%, shrimp 1.6%, rate 1.6%, popul 1.6%) – focuses on effects of diet and food on species growth, especially larvae and shrimp.

(70) Cluster 29) (sediment 8.6%, california 5.1%, gulf 3.1%, river 2.8%, water 2.0%, pacif 1.6%, reef 1.5%, gulf.california 1.5%, baja 1.5%, basin 1.1%) – focuses on sediments, especially in the California Gulf and river waters.

(54) Cluster 30) (equat 18.3%, solut 8.2%, numer 2.6%, solv 2.2%, mixtur 1.1%, channel 1.0%, closur 0.9%, comput 0.8%, hard 0.8%, numer.solut 0.8%) - focuses on numerical solutions of equations.

(90) Cluster 31) (control 16.9%, stabil 4.9%, system 3.6%, feedback 3.2%, synchron 3.0%, robot 2.3%, output 2.0%, chaotic 1.6%, dynam 1.2%, paper 1.2%) - focuses on feedback control and stabilization of systems, especially robots and chaotic dynamical systems.

(52) Cluster 32) (plant 18.0%, shoot 4.2%, seedl 2.2%, tree 2.1%, leaf 2.0%, season 1.9%, leav 1.7%, liana 1.6%, stem 1.4%, growth 1.3%) – focuses on plants, emphasizing shoots, seedlings, tree leafs and growth phenomena.

(53) Cluster 33) (asphalten 10.5%, polym 5.3%, oil 3.7%, polymer 3.3%, chitosan 2.6%, copolym 1.8%, graft 1.7%, particl 1.6%, monom 1.6%, blend 1.4%) – focuses on asphaltenes in oil, and graft copolymers.

(72) Cluster 34) (degreesc 10.3%, temperatur 4.8%, coat 3.2%, powder 3.1%, gel 2.3%, glass 2.1%, corros 2.0%, sol 1.8%, phase 1.5%, composit 1.3%) – focuses on temperature and corrosion effects of powder coatings and gel coatings.

(82) Cluster 35) (protein 27.7%, kda 5.4%, bind 2.2%, activ 1.7%, subunit 1.3%, toxin 1.2%, beta 1.1%, amino 1.0%, termin 1.0%, membran 1.0%) – focuses on large mass proteins, especially binding activity of toxins.

(113) Cluster 36) (star 11.5%, emiss 7.1%, jet 3.9%, veloc 2.6%, observ 2.5%, nebula 2.5%, radio 2.0%, maser 2.0%, line 1.9%, sourc 1.9%) – focuses on velocity of jet emissions from stars.

(132) Cluster 37) (cell 49.5%, express 1.4%, cultur 1.3%, dna 1.2%, apoptosi 0.9%, cell.line 0.9%, prolifer 0.7%, lymphocyt 0.6%, receptor 0.6%, secret 0.5%) – focuses on DNA analysis of cell cultures.

(58) Cluster 38) (popul 12.4%, genet 5.6%, forest 4.4%, allel 3.0%, stock 1.3%, divers 1.2%, mexican 1.1%, area 1.1%, african 1.1%, mexico 1.1%) – focuses on forest population genetics.

(76) Cluster 39) (quantum 20.2%, state 2.9%, hamiltonian 2.7%, motion 2.0%, particl 1.5%, classic 1.4%, reson 1.3%, quantum.mechan 1.2%, potenti 1.2%, system 1.0%) – focuses on quantum states, especially hamiltonians and associated particle motions.

(88) Cluster 40) (compound 11.6%, nmr 9.4%, reaction 3.6%, structur 2.0%, methyl 1.9%, deriv 1.7%, rai 1.6%, synthesi 1.2%, substitut 0.9%, spectroscopi 0.9%) – focuses on NMR analysis of compounds formed by reactions.

(176) Cluster 41) (patient 48.6%, diseas 1.6%, group 1.4%, month 1.3%, infect 1.1%, year 1.0%, treatment 0.9%, surgeri 0.6%, surgic 0.5%, mean 0.4%) – focuses on treatment of patients, especially with infectious diseases.

(79) Cluster 42) (puls 11.1%, optic 8.2%, laser 5.8%, fiber 5.7%, power 3.5%, pump 1.9%, polar 1.8%, birefring 0.9%, measur 0.7%, interferomet 0.6%) – focuses on pulsed fiber optic lasers.

(85) Cluster 43) (rat 27.7%, liver 3.5%, dai 2.3%, activ 1.4%, induc 1.1%, administr 1.0%, increas 1.0%, dose 0.9%, anim 0.8%, group 0.8%) –focuses on liver damage in rats.

(51) Cluster 44) (slip 5.2%, veloc 2.4%, fault 2.1%, volcan 1.5%, wind 1.3%, flow 1.3%, uncertainti 1.3%, estim 1.1%, surfac 1.1%, vertic 1.1%) – focuses on geology, emphasizing earthquakes and volcanic action.

(70) Cluster 45) (equat 9.1%, field 8.1%, fermion 3.3%, solut 2.2%, symmetri 2.0%, string 1.6%, wave 1.4%, tensor 1.2%, cmm 0.9%, boson 0.9%) – focuses on solutions to field equations, including bosonic and fermionic symmetric nuclei.

(75) Cluster 46) (season 6.1%, fish 4.6%, femal 3.8%, abund 3.0%, year 1.9%, male 1.8%, california 1.5%, size 1.4%, fisheri 1.3%, biomass 1.3%) – focuses on fish abundance as a function of seasonal and ecological variables.

(57) Cluster 47) (blood 8.5%, lead 3.7%, group 3.3%, cow 3.3%, serum 2.4%, test 1.9%, blood.lead 1.8%, vaccin 1.6%, sampl 1.4%, anim 1.3%) - focuses on epidemiological studies of health problems in humans and animals, emphasizing blood lead levels in different groups.

(91) Cluster 48) (atom 4.6%, calcul 3.6%, bond 2.5%, structur 2.5%, molecul 2.0%, electron 2.0%, energi 2.0%, reaction 1.4%, initio 1.3%, kcal 1.2%) - focuses on calculations of atomic and molecular structures, emphasizing bond structures and electron energies.

(64) Cluster 49) (electrod 5.0%, reaction 4.4%, metal 3.6%, oxid 3.3%, electrochem 3.0%, iron 1.7%, slag 1.7%, discharg 1.4%, hydrogen 1.2%, carbon 1.1%) – focuses on oxidation reactions at metal electrodes in electrochemical systems.

(104) Cluster 50) (speci 35.5%, forest 2.0%, habitat 1.7%, bat 1.6%, tree 1.3%, plant 1.2%, divers 1.1%, ant 1.1%, type 0.9%, nest 0.8%) – focuses on species populations and evolution, emphasizing smaller species in forest habitats.

(77) Cluster 51) (neuron 5.1%, sleep 3.7%, rat 3.2%, dose 2.1%, induc 1.7%, antinocicept 1.6%, receptor 1.2%, activ 1.2%, increas 1.1%, administr 1.0%) – focuses on the role of neurons in the sleep cycle, emphasizing drug tests in rats.

(83) Cluster 52) (alloi 11.3%, phase 6.2%, particl 2.5%, materi 2.5%, electron 2.0%, composit 1.8%, grain 1.4%, microscopi 1.4%, size 1.2%, temperatur 1.1%) – focuses on material phases in alloys, including particle properties.

(62) Cluster 53) (children 15.6%, ag 2.3%, nurs 2.2%, group 1.6%, diarrhea 1.2%, test 1.1%, drug 1.1%, year 1.0%, treatment 1.0%, lamb 0.8%) - focuses on children’s illnesses, especially those with intenstinal symptoms such as diarrhea.

(62) Cluster 54) (theori 4.8%, function 3.6%, densiti 2.0%, entropi 2.0%, properti 1.2%, energi 1.2%, densiti.function 1.1%, salt 1.0%, calcul 1.0%, walk 1.0%) – focuses on theoretical density functional procedures for computing system energetics and entropies.

(80) Cluster 55) (activ 15.6%, enzym 4.8%, compound 4.0%, extract 3.7%, beta 3.6%, inhibit 2.3%, isol 1.5%, mug 1.4%, alpha 1.3%, inhibitor 0.7%) –focuses on enzyme activity for identifying compounds in extracts.

(66) Cluster 56) (optim 7.9%, design 3.6%, paper 2.1%, system 2.1%, methodolog 1.9%, semant 1.7%, reward 1.4%, base 1.3%, learn 1.0%, term 1.0%) – focuses on semantic analyses to optimize the design of document constructs.

(98) Cluster 57) (space 21.2%, prove 2.4%, point 1.7%, manifold 1.3%, set 1.3%, compact 1.3%, connect 1.1%, space.time 1.0%, theorem 0.9%, map 0.9%) – focuses on mathematical systems in different spaces, emphasizing proofs of theorems.

(72) Cluster 58) (ion 4.0%, adsorpt 3.6%, water 3.5%, solut 2.8%, aqueou 2.0%, extract 1.7%, rock 1.5%, phase 1.0%, remov 1.0%, miner 1.0%) – focuses on absorption of ions from aqueous solutions.

(90) Cluster 59) (health 4.7%, citi 4.5%, mexico 3.5%, water 2.4%, radon 2.1%, research 2.0%, mexico.citi 1.7%, land 1.3%, countri 1.3%, sampl 1.1%) – focuses on health impacts of radon and water, especially in Mexico City.

(93) Cluster 60) (patient 10.9%, case 4.8%, syndrom 2.5%, chromosom 2.2%, prl 2.0%, diagnosi 1.7%, diseas 1.7%, allel 1.3%, lesion 1.1%, year 1.1%) – focuses on clinical studies of patients with different medical problems.

(101) Cluster 61) (algorithm 11.0%, method 4.9%, estim 2.1%, network 1.9%, imag 1.7%, techniqu 1.4%, power 1.3%, base 1.3%, error 1.0%, filter 1.0%) – focuses on algorithms for analyzing different physical systems.

(83) Cluster 62) (acid 7.2%, cultur 5.6%, medium 2.5%, product 1.9%, glucos 1.7%, lectin 1.7%, gonad 1.6%, yeast 1.4%, concentr 1.2%, ferment 1.2%) – focuses on acid content of culture mediums.

(92) Cluster 63) (model 19.7%, dynam 3.6%, time 1.4%, predict 1.3%, simul 1.3%, system 1.0%, experiment 0.9%, frame 0.8%, soil 0.7%, data 0.7%) – focuses on predictive temporal dynamical models.

## APPENDIX 9 – DATA COMPRESSION CLUSTERING METHOD

In order to compare the similarities between the abstract, we have used a compression algorithm (44). Recently a zipping method to recognize the subject treated in a text was proposed (44). This method uses that the entropy of a string can be measured when this string is zipped (compressed). The main idea is that when one compresses two strings, one after another one, the compression rate will increase if the second string is similar to the first one, and then the zipped string will have less disorder (entropy) than the previous two strings. Then, this algorithm considers that two close papers will have a relative informational entropy close to zero. This algorithm is based on the statistical basis, this means that it works better with long files. In our case the abstracts are not actually long files, however this method works properly as will see later

First we look for the repeated authors and analyze the corresponding abstracts in order to find close related abstracts. Then we applied the entropic algorithm based on zipping (compressing) files using the following formula:

Entropy = (Length(zip(a+b))-Length(zip(a)) - Length(zip(b+b))+Length(zip(b)) )/ Length(b).

Where a and b are the abstracts to be analyzed, zip indicates the zipped abstract.

We have considered that the candidates to be repeated abstracts present an entropy less than 0.20; Because a Gaussian distribution can be fitted to the entropy distribution as can be seen in fig. A. and the real entropy distribution separates from this Gaussian distribution a value close to 0.20 indicating that the similarities of these abstracts are not produced by random conditions. The zero distance is for identical abstract in both directions, also you can find that some zero entropic distance can appear in some abstracts contained in others.

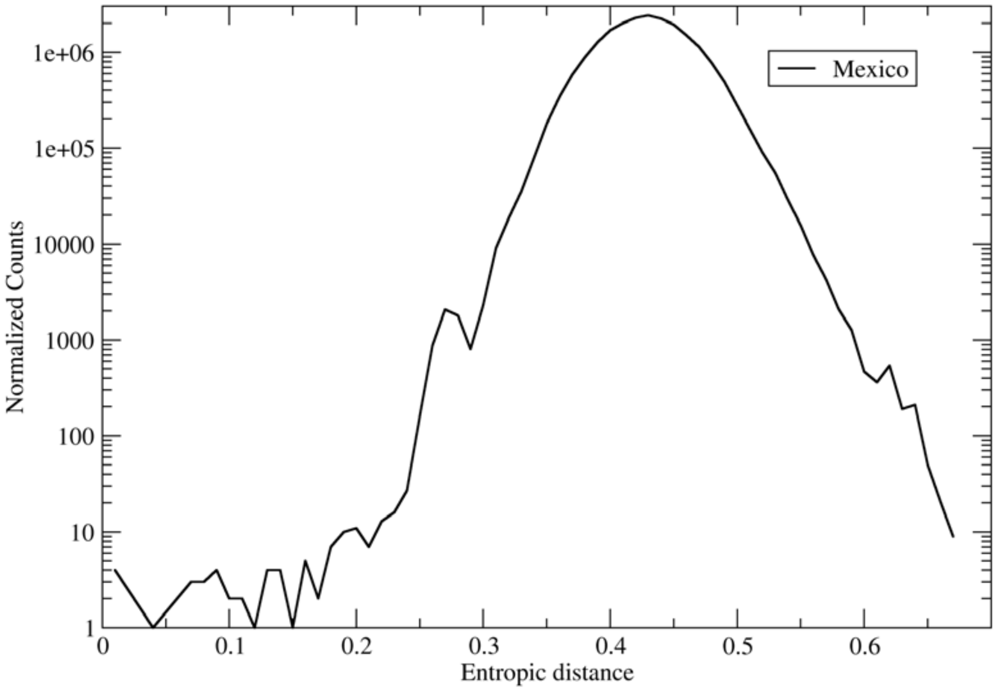


Figure A. The normalized distribution of the entropic distance between pairs of papers (integral equal to one).

With this method we have detected around of 1% of similar papers appearing in the 2001 Mexico scientific papers. This rate is of the same order to other countries or thematic similarities results.

One important character of the entropy distribution that deserves explanation is the kink around 0.26. This kink appears clearly in Mexico analysis. We guess that this kink is due to the fact that there is a 'natural' clustering (areas, categories, scheme classification, etc.) where most of the papers can be classified. We have tried several algorithms, to determine with this distance the clusters or classification groups.

Our fundamental objective is to automate the classification of records into pre-defined categories for example DTIC theme. Our concept is to compare the complete abstract of each record against lexicons for each pre-determined DTIC theme and then assign each record to the category that provides the best match. In other form, we are defining the closest category to the abstract and also we can give the following closer categories.

In this form, we have assigned, in an computerized form, first level DTIC theme for each paper in the Mexico core competencies analysis. We think that our approach has some range of error. The main limitation in this approach is our form to define a good dictionary for each DTIC theme. The algorithm we use requires a good definition of dictionaries for each DTIC theme.

First we have defined 19 patron text or lexicons for 19 DTIC themes. With these 19 DTIC theme dictionaries we have compressed the 4529 abstracts and selected, using the best compression rate, the corresponding first level categorization theme for each abstract. The form to define the patron text and lexicon is explained in Appendix 9.
Moreover to Entropy formula, we use two other variants as:

B)
Entropy = (Length(zipL(A+b))-Length(zipL(A)) - Length(zipL(b+b))+Length(zipL(b)) )/ Length(b).

where zipL indicates a zipping process with the lexicon as parameter. A variant allowing calculate in less time.

C)
Entropy = (Length(zipL(L+b))-Length(zipL(L)) - Length(zipL(b+b))+Length(zipL(b)) )/ Length(b).

where the difference is that we have used the Lexicon as a patron text.

We start with a lexicon for each theme and save the Entropy for each lexicon, we repeat this procedure with the resting lexicons and select the lowest entropy between all possibles lexicons, this is the best fit and write the corresponding theme. Then this procedure is repeated with the following abstracts. The complete procedure takes less than 10 minutes in a Pentium 4 PC with Linux. The computational time is of the order of 6 to 3 hrs from the A to C entropy measurement.

All this procedure depends strongly on the form we define the dictionaries.

In this analysis the pragmatical form to define good dictionaries was the following. We selected more than 250 abstracts of papers appearing during 2001 in scientific journals that we considered are in the specific DTIC theme (the list of journals appears in the Tables in the attached spreadsheet). This selection was a manual form to define DTIC theme to the journals. With these 250-2000 different abstracts from different journals in each first-level classification, we proceed to perform a keyword detection in each theme. In order to do this, we follow the procedure indicated in reference 54. This method uses the standard deviation of the distance between successive occurrences of a word in a text as an indicator of the relevant words. The standard deviation is actually close to the entropy (55), We follow this method and we select the words with normalized standard deviation of the distance between successive occurrences higher than one as words in the corresponding dictionary. The selection of this standard deviation is due to the fact that standard deviation less than one indicates a random distribution of the words around the text. In this way we have defined the relevant words in each DTIC first level theme in 2001. Thus, we generate the lexicons for each DTIC theme by choosing technical journals that reflect the category of interest closely, selecting records from those journals, and extracting the word patterns from those records.

The difficulty in this method is to get the complete set of abstracts, We expended several weeks in order to download the information from ISI web-site. But when we have the text of the corresponding sets of abstracts the computing time is around 5 min. per theme. Of course we can use books in order to obtain the relevant words in each theme.

Here it is important to note that with this method it is possible to analyze all the abstracts. The results are the following for automated classification with relative entropy defined by A), B) and C) is given in tables

Automated Classification A Formula

| | |
|---|---|
| Physics | 23% |
| Biological and Medical sciences | 32% |
| Chemistry | 8% |
| Agriculture | 8% |
| Mathematical and Computer sciences | 25% |
| Earth sciences and Oceanography | 7% |
| Material sciences | 12% |

Automated Classification B Formula

| | |
|---|---|
| Physics | 23% |
| Biological and Medical sciences | 54% |
| Chemistry | 8% |
| Agriculture | 10% |
| Mathematical and Computer sciences | 15% |
| Earth sciences and Oceanography | 6% |
| Material sciences | 26% |

Automated Classification C Formula

| | |
|---|---|
| Physics | 16% |
| Biological and Medical sciences | 38% |
| Chemistry | 6% |
| Agriculture | 7% |

| | |
|---|---|
| Mathematical and Computer sciences | 11% |
| Earth sciences and Oceanography | 4% |
| Material sciences | 18% |

## APPENDIX 10 – LEXIMANCER CLUSTERING METHOD

Descriptions of the Leximancer algorithms have been published elsewhere [A10-1]. However, mention will be made here of particular settings employed in this case.

The abstracts were pre-processed by inserting the title and the content of each into its own file. No account was taken of the other metadata for this analysis.

The Leximancer system offers several possible strategies for extraction of concept seeds, depending on the nature and requirements of the final application. For content analysis purposes, stylistic, cultural and emotive concepts may be of interest. It was judged that for the purposes of this investigation, factual and specific semantic categories are preferred. The strategy used to achieve this will be described here briefly for reference purposes: To avoid stylistic concepts which are common to scientific abstracts, 250 ranked top-level concepts were extracted where each concept must be positively relevant to at least one of the abstracts – this removes general concepts which are ubiquitous within the data. These were then tidied manually to remove a few important but rather non-specific ideas such as 'length', 'compared', 'number', 'high', 'low'. To add a broad second level of more specific concepts, 500 ranked highly specific concepts were extracted. Bi-gram concept selection was enabled to augment this set with common bi-grams. This set of approximately 750 seeded concepts was then trained into a thesaurus, to form a flat list of lexical representations for each. The selected number of concepts is fairly large, and was guided by the large conceptual diversity of the data. Inspection of the final concept frequency histogram and classification coverage indicate that coverage of the material by the thesaurus is adequate.

The thesaurus was then used to classify the material at a resolution of 2 sentences. The normal segment length for prose is 3 sentences, but abstracts display a much more condensed discourse. This can be measured by optimising the machine learning parameters. This tuning process showed that the average conceptual coherence length was closer to two sentences in this case.

A10-1 Smith, A.E. Automatic Extraction of Semantic Networks from Text using Leximancer. HLT-NAACL 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics - Companion Volume. ACL, May 2003. Demo23-Demo24.